Best practice guidance for linear mixed-effects models in psychological science

Best practice guidance for linear mixed-effects models in psychological science

Lotte Meteyard[1*]

Robert A.I. Davies[2]

Word count: 18,952 (including tables)

Keywords: linear mixed effects models, hierarchical models, multilevel models.

Affiliations

[1*] *Corresponding author*
School of Psychology & Clinical Language Sciences,
University of Reading, Berkshire, UK, RG6 6AL.
l.meteyard@reading.ac.uk

[2]
Department of Psychology,
Lancaster University, Lancaster, UK, LA1 4YF.

Acknowledgements

This work was supported by a British Academy Skill Acquisition Grant SQ120069 and University of Reading 2020 Research Fellowship to LM. We would like to thank all the students, researchers and colleagues who have discussed mixed models with the authors over the last few years, and those who responded to the survey. Our thanks go also to Elizabeth Volke who helped LM prepare the survey and collate results whilst visiting the School of Psychology & Clinical Language Sciences, University of Reading, on an Erasmus studentship. LM would like to thank Morgan and Marcella Meteyard Whalley, for enabling the time and mental space to get this project started and finished. We are grateful to reviewers and colleagues whose comments improved the manuscript considerably.

Abstract

The use of Linear Mixed-effects Models (LMMs) is set to dominate statistical analyses in psychological science and may become the default approach to analyzing quantitative data. The rapid growth in adoption of LMMs has been matched by a proliferation of differences in practice. Unless this diversity is recognized, and checked, the field shall reap enormous difficulties in the future when attempts are made to consolidate or synthesize research findings. Here we examine this diversity using two methods – a survey of researchers (n=163) and a quasi-systematic review of papers using LMMs (n=400). The survey reveals substantive concerns among psychologists using or planning to use LMMs and an absence of agreed standards. The review of papers complements the survey, showing variation in how the models are built, how effects are evaluated and, most worryingly, how models are reported. Using these data as our departure point, we present a set of best practice guidance, focusing on the reporting of LMMs. It is the authors' intention that the paper supports a step-change in the reporting of LMMs across the psychological sciences, preventing a trajectory in which findings reported today cannot be transparently understood and used tomorrow.

1.0 Introduction

Linear Mixed-effects Models (LMMs) have become increasingly popular as a data analysis method in the psychological sciences. They are also known as hierarchical or multilevel or random effects models (Snijders & Bosker, 2011). LMMs are warranted when data are collected according to a multi-stage sampling or repeated measures design. That is, when there are likely to be correlations across the conditions of an experiment because the conditions include the same participants or participants who have some association with each other. Multi-stage sampling can arise naturally when collecting data about the behavior or attributes of participants recruited, e.g., as students from a sample of classes in a sample of schools, or as patients from a sample of clinics in a sample of regions. Repeated measures occur when participants experience all or more than one of the manipulated experimental conditions, or when all participants are presented with all stimuli. Such investigations are common in psychology. These designs yield data-sets that have a *multilevel or hierarchical* structure. Participant-level observations, e.g., an individual's measured skill level or score, can be grouped within the classes or schools, clinics or regions from which the participants are recruited. Trial-level observations, e.g., the latency of response to a stimulus word, can be grouped by the participants tested or by the stimuli presented (Baayen, Davidson, & Bates, 2008). We expect that the responses made by a participant to some stimuli will be correlated, or that responses from children in the same class or school or region will be correlated, or that responses to the same stimulus item across participants will be correlated. The hierarchical structure in the data (the ways in which data can be grouped) is associated with a hierarchical structure in the error variance. LMMs allow this structure to be explicitly modelled.

We review current practice for LMMs in the psychological sciences. To begin, we present an example of a mixed-effects analysis (Section 1.1), with the aim of

clearly illustrating how random effects relate to fixed effects. Researchers who are comfortable in their conceptual understanding of LMMs may wish to skip this part. Following the example, we present data from a survey of researchers (Section 2.0) and a review of reporting practices in papers published between 2013 and 2016 (Section 3.0). Our observations reveal significant concerns in the community over the implementation of LMMs, and a worrying range of reporting practices in published papers (Section 4.0). Using the available literature, we then present best practice guidance (Section 4.1) with a bullet-point summary (Section 5.0). To preempt two key conclusions, researchers should be reassured that there is no single correct way to implement an LMM, and that the choices they make during analysis will comprise one path, however justified, amongst multiple alternatives. This being so, to ensure the future utility of our findings, the community *must* adopt a standard format for reporting complete model outputs (see the example tables in Appendix 5). All appendices and data are available at osf.io/bfq39.

1.1 An example

Our example is introductory but it is not intended as a step-by-step tutorial. We provide an explanation of mixed-effects models without recourse to algebra or formulae. In particular, we discuss random intercepts and random slopes in the context of this example, and how these can be fit alone (intercepts or slopes only) or together (intercepts and slopes) for a given fixed effect predictor. In our experience as researchers and teachers, this is the biggest conceptual hurdle to understanding and working with LMMs.

A subset of data from Meteyard and Bose (2018) has been used, and scripts and data are available from osf.io/bfq39 – Files – LMMs_BestPractice_Example.R and NamingData.txt for those wishing to recreate the analysis and graphs[1]. For those

[1] We have left annotations and comments between the two authors in the script, to illustrate the work-in-progress nature of a coding script

wishing to see the model output, osf.io/bfq39 – Files – LMMs_BestPractice_Example_withOutput is available as both an R script and a text file.

To collect the data, ten individuals with aphasia completed a picture naming task. Stimuli comprised 175 pictures from the Philadelphia Naming Test (Roach et al., 1996). The experiment tested how cues presented with the pictures affected naming accuracy, and each picture was presented with four different cues. Thus, each participant was presented with each picture four times, and the study conformed to a repeated measures design. The four cues were: a word associated to the naming target (towel - bath); an unassociated word (towel - rice); the phonological onset (towel – 't'); and a tone. Given previous findings, we predicted that a phonological onset cue or an associated word cue would improve naming accuracy, relative to an unassociated word cue or the tone cue. The experiment also tested how the properties of the target name affected naming accuracy. Here we will look at the length of the word (in phonemes) and the frequency of the word (log lemma frequency). We predicted that words with more phonemes (longer words) would be harder to name, as reflected in reduced response accuracy, whereas words with higher frequency would be easier to name, as seen in increased accuracy.

In conventional mixed-effects modeling terms, given this design, we have three fixed effects. Cue type is a factor with four levels (the different cues). Length and frequency are two continuous predictors that have a value associated to each target picture name. The random effects are associated with the unexplained[2] differences between the participants (10 participants, each of whom completed 700 trials) and the items (175 items, each associated with 40 observed responses). Participants and items were sampled from respective person or picture populations:

[2] We do not have a complete explanation for why different participants or items are associated with variation in responses, so there will always be some error variance associated with participants and items that is 'unexplained'. Any explanations or predictions that we do have can be included in the model as fixed effect predictors.

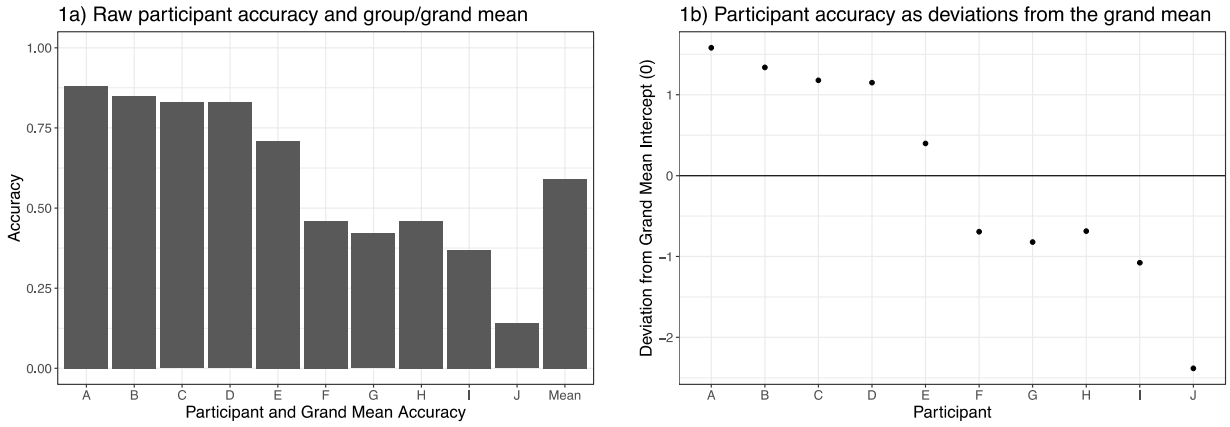
each participant and each item can be seen to be a *sampling unit* from a wider population. Intuitively, responses by a participant will tend to be correlated because one person may be more or less accurate than another, on average. Responses to each item will tend to be correlated because one picture will be more or less difficult than another. For simplicity, we are going to illustrate random effects for participants only. Graphs are generated from mixed-effects models with all fixed effects predictors but *only* the random effect under consideration (see Figures 1-4). This is so we can consider each case in isolation.

The simplest possible random effect to include in the mixed-effects model would be the random effect of participant on intercepts, in an intercepts only model. What does that mean? To start, we can calculate the average accuracy (grand mean) across all participants' responses. However, the participants differ in the severity of their aphasia, and this variation leads to differences between participants in their average naming accuracy (Figure 1a). To account for this, we can model the random variance in intercepts due to unexplained differences between participants: the random intercepts by participants. Figure 1a shows the raw data, with each participant's accuracy (averaged across all the trials they completed) and the grand mean. It is clear that some participants are above the mean and some are below it. Because we are modelling how each participant deviates from the grand mean, it is convenient to scale the units for these differences as *standard deviations* from the grand mean, centered at zero. Figure 1b shows the random intercepts for participants (extracted from a mixed-effects model that included the fixed effects plus just the random intercepts by participant) where zero represents the grand mean. This plot shows the *difference* between each participant's accuracy and the grand mean accuracy. The model output tells us that the variance associated with random intercepts is 1.58 (SD=1.257). So, on average, participant-level intercepts vary around the grand mean by 1.257 SD units. Given that the units for measurement of

accuracy go from 0 to 1, we can interpret this as quite a large amount of variation across participants. This is clearly illustrated in Figure 1a and 1b.

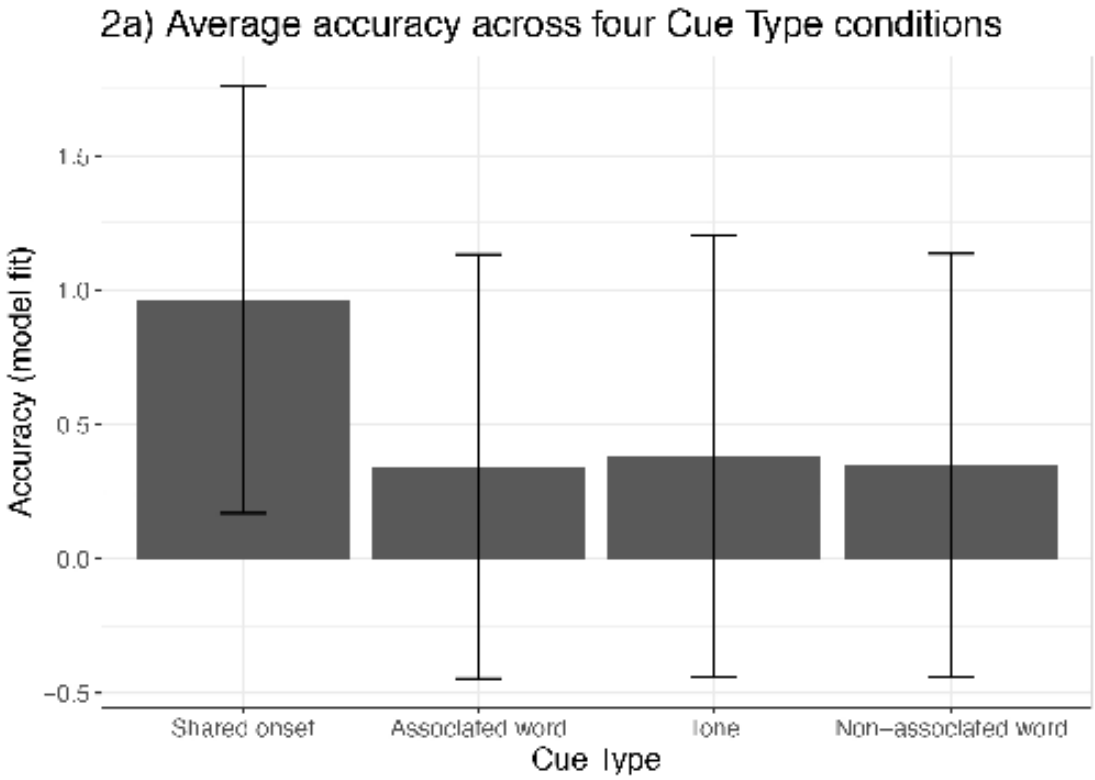
----- FIGURE 1 -----

Figure 1 Title: Illustrations for Participant Intercepts for Naming Accuracy



Our readers will know that participants may differ not only in the level of performance (average accuracy of response) but also in the ways in which they are influenced by the experimental conditions or by the stimulus attributes. We can account for random differences between sampled participants in their response to cue type by specifying a model term corresponding to *random slopes* for the effect of cue type, that is, to deviations between participants in the coefficient for cue type. We can calculate the average naming accuracy *within each cue condition* across participants. To get the fixed effect result, we can then (as in an ANOVA) compare the four cue types to each other and see *on average* the effect of cue type on naming accuracy. Figure 2a shows the average response accuracy per condition, illustrating this fixed effect for cue type.

--- Figure 2a here –



The trends in the plot suggest that cues which share the target onset (known as a 'phonological cue' in aphasia research) increase accuracy relative to the other three cue types. When we model random slopes for cue type over participants (i.e. slopes only, without random intercepts), we aim to gauge how the effect of cue type differs across participants. In this experiment, cue type is a factor with four levels, so we are concerned with the variation among participants in how the average accuracy of response differs between the four conditions.

-- Figure 2b here –

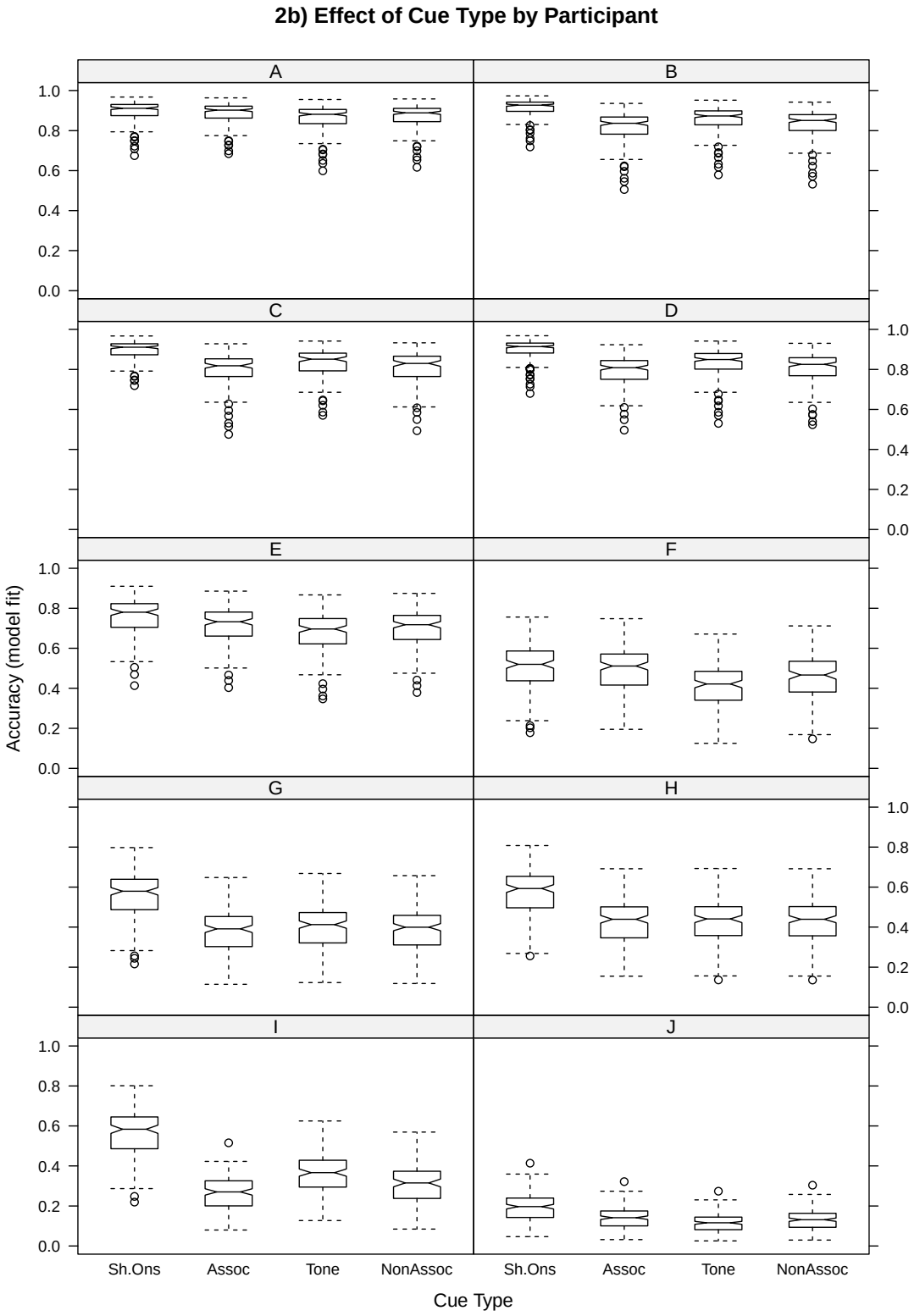
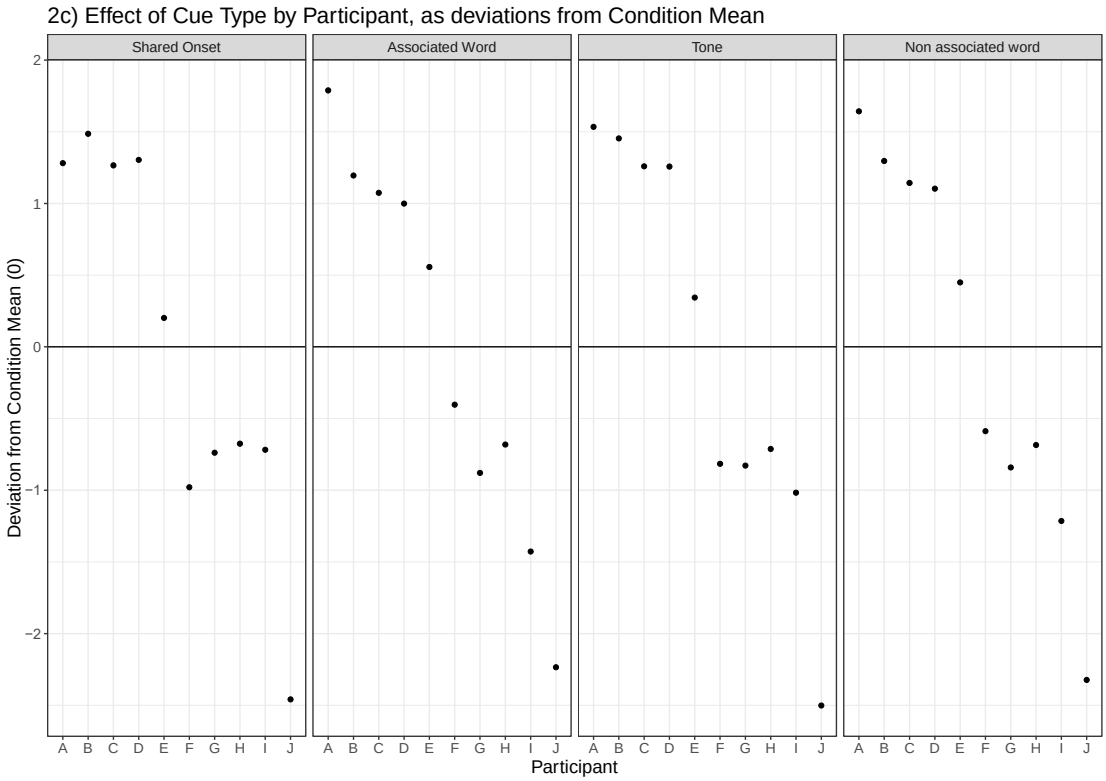


Figure 2b shows the individual participant data for each condition. It is clear that not all participants show the same effect of phonological cueing. For example,

participants who are highly accurate across all conditions (Participants A to D) show ceiling effects, so there is not much scope for phonological cueing to improve naming further. So, what is the spread (variance) of deviations between participants around the average effect of cue type? Figure 2c shows the participant random slopes estimated for the effect of cue type. This shows how *within each cue type condition* the effect for different participants varies around the mean accuracy of responses under that condition.

-- Figure 2c here –



The model output tells us the variance in slopes associated with each cue type (Shared onset SD = 1.268, Associated word SD = 1.259, Tone SD = 1.310 and Non-associated word SD = 1.254). So, on average, within each condition, participants vary around the mean by ~1.3 units. The model output also tells us how the by-subjects deviations in the slopes of the effects of cue type conditions are

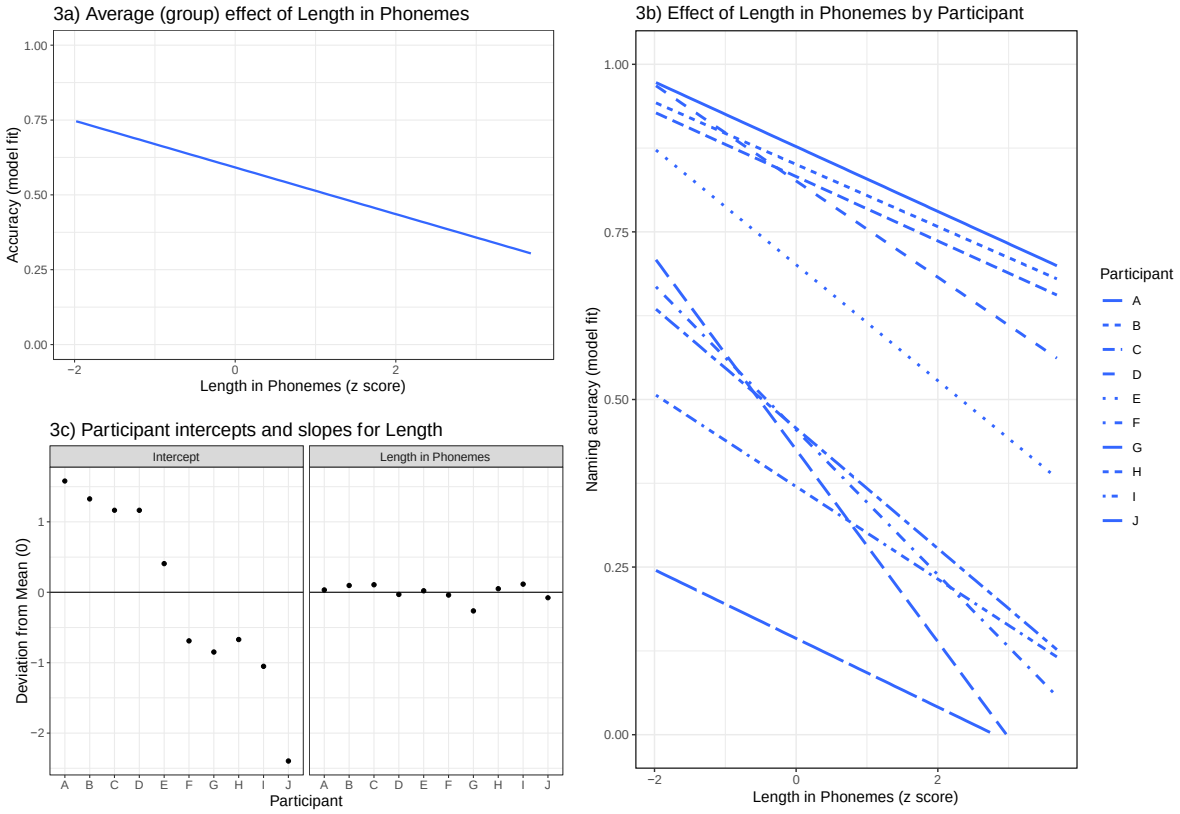
correlated with each other, with high positive correlations (0.94 - 0.99). A per-subject deviation in response to one condition will, predictably, correlate with the deviation for the same participant in response to other conditions. This is perhaps unsurprising given how much the participants vary between each other, a variation that is driven principally by the severity of their aphasia. Put another way, the main explanation of participants' performance *across* the different cue conditions comes from accounting for the differences between the participants. This is a nice example of the *variance-covariance* structure in the data – i.e. where variation arises and how it is related across groupings in the data.

For the continuous fixed effect predictors, the term 'random slope' will make more intuitive sense, and here we will model both random intercepts and random slopes for the effect of length across participants. For a more complete account of the data, we will also ask the model to fit the *covariance* for intercepts and slopes – that is, to model them as *correlated*. For example, participants who are more accurate (higher intercept) may show a stronger effect of length (steeper slope), resulting in a positive correlation between intercept and slope. First, to see how word length affects naming accuracy, we look at the slope of naming accuracy when we plot it against length, illustrating the average effect of length (see Figure 3a). By fitting random intercepts *and* random slopes for word length over participants, we model both the differences between participants in overall accuracy (see Figure 1) *and* the between-participant differences in the slope for the effect of length. To illustrate this, we have plotted the fitted values from a model with random intercepts and with random slopes for word length over participants. Figure 3b shows the separate estimated slope of the length effect for each participant. More accurate participants have higher intercepts, and participants show differences in how steep or shallow the slope for length is. Steeper slopes mean a stronger effect of length on naming accuracy. Finally, we plotted the same data as the random effects – that is, the per-subject deviations from the average intercept and from the average slope

(Figure 3c). From this plot, we can see that deviation in overall accuracy (i.e. random variation in intercepts by participant) is much greater than in slopes for length.

----- FIGURE 3 -----

Figure 3 Title: Illustrations for Participant Intercepts and Slopes for Length in Phonemes



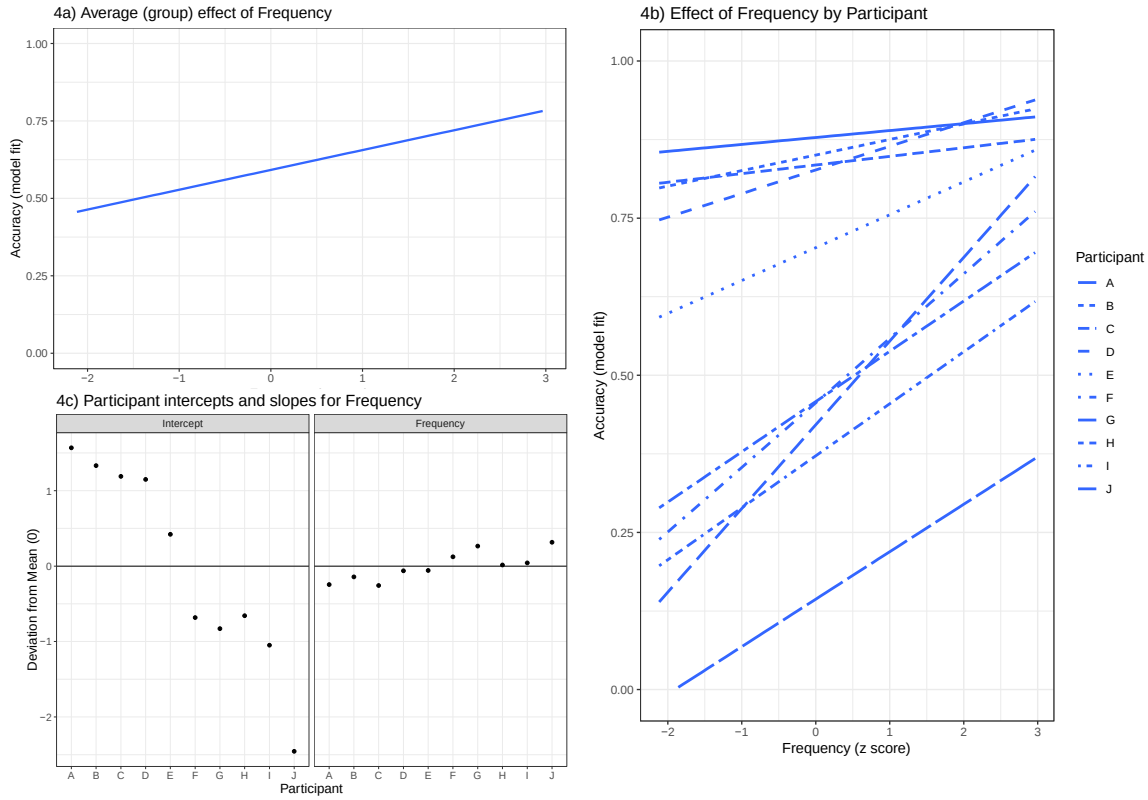
The model output gives us the variance associated with participant intercepts (SD = 1.259), with values consistent with those we have seen previously for participant intercepts. It also gives us the variance associated with the effect of Length in Phonemes, with an SD = 0.135. This is much smaller than the variation associated with participant intercepts, and that tells us that (as seen in Figure 3c) participants show much more variation in the overall accuracy of their naming than they do in how their naming accuracy is affected by the lengths of the words they

name. The correlation (covariance) between the slopes for Length and the participant intercepts is positive and relatively low (0.34). Thus, there is some tendency for participants with higher accuracy (higher intercepts) to show greater effects of Length (steeper slopes).

Finally, the same process can be applied to the effect of target name frequency, and this is illustrated by Figures 4a-c. Here we can see more variation between participants in how frequency affects naming accuracy.

----- FIGURE 4 -----

Figure 4 Title: Illustrations for Participant Intercepts and Slopes for Frequency



The model output gives us the variance associated with participant intercepts (SD = 1.268), again consistent with previous estimates. It also gives us the variance associated with the effect of Frequency, with an SD = 0.200. This is larger than the variation associated with the effect of Length, as illustrated in the spread of dots (per-

subject deviations) in Figure 4c. The correlation between the random slopes for the effect of Frequency and the participant random intercepts is large and negative (-0.84). Participants with higher accuracy (higher intercepts) show a reduced effect of Frequency (shallower slopes), this is clearly reflected in Figure 4b.

We hope that this example has done two things. First, clearly explained the concept of random and fixed effects. Second, highlighted just how informative the random effects can be.

1.2 The ascendency of mixed models

LMMs have grown very popular in modern Psychology because they enable researchers to estimate (fixed) effects while properly taking into account the random variance associated with participant, items or other sampling units. From under 100 Pubmed citations in 2003, the number of articles referring to LMMs rose to just under 700 by 2013 (see Figure 5), the starting year in our review of LMM practice. This popularity is associated with an increasing awareness of the need to use LMMs. However, the growth in popularity has been associated with a diversity among approaches that will incubate future difficulties. In simple terms, variation in current reporting practices will make meta-analysis or systematic review of findings near impossible. The present article examines the diversity in modeling practice and outlines the features of a reproducible approach in using and reporting mixed-effects models.

Historically, the dominant approach for repeated measures data in psychology has been to aggregate the observations. Typically, in Psycholinguistics, a researcher would calculate the mean latency of response for each participant, by averaging over the RTs of each stimulus, to get the average RT by-participants within a condition for a set of stimuli (e.g., per cue type, if our example were a naming latency study). In a complementary fashion, mean RTs for each stimulus

would be calculated by averaging over the RTs of each participant, to get the average RT by-items within a condition (e.g., each cue type condition). The means of the by-participants or by-items latencies would be compared using Analysis of Variance (ANOVA) in, respectively, by-participants (F1 or F_s) or by-items (F2 or F_i) analyses. If s/he was seeking to correlate the average latency of responses by-items with variables indexing stimulus properties, or by-participants with variables indexing participant attributes, s/he would use multiple regression to estimate the effects of item or participant attributes on the averaged latencies. A series of analyses dating back over 50 years have shown that these approaches suffer important limitations (Baayen et al., 2008; Clark, 1973; Coleman, 1964; Raaijmakers Schrijnemakers, & Gremmen, 1999).

As Clark (1973; after Coleman, 1964) noted, researchers seeking to estimate experimental effects must do so in analyses that account for random differences in outcome values both between participants and between items. The random differences can include by-participants or by-items deviations from the average outcome (e.g., fast or slow responding participants, see Figure 1), or from the average slopes of the experimental effects (e.g., individual differences in the strength of an experimental effect, see Figures 2-4). The presence of random differences in the intercept or in the slope of the experimental effect *between-items* meant, Clark (1973) observed, that the at-the-time common practice of using only by-subjects' ANOVAs to test differences between conditions in mean outcomes was likely to be associated with an increased risk of committing a Type I error. Such errors arise in Null Hypothesis Significance Testing (NHST) where the researcher calculates a test statistic (e.g., t corresponding to a difference between conditions) and compares its value with a distribution of hypothetical test statistics generated under the *null hypothesis* assumption of no difference. A p value indicates the proportion of test statistic values, in the null hypothesis distribution, equal to or more extreme than the test statistic calculated given the study data (Cassidy, Dimova, Giguère, Spence, &

Stanley, 2019). Errors arise when a researcher rejects the null hypothesis when there is no substantial underlying difference in outcomes between conditions. Ignoring random variation in outcomes among stimulus items can mean that significant effects are observed and interpreted as experimental effects, when they are in fact due to uncontrolled variation amongst items (e.g., effects seen in by-participant average RTs are in fact driven by a 'fast' or 'slow' item influencing the means). This was termed the language-as-fixed-effect fallacy.

Clark's (1973) remedy was to calculate F1 and F2 and then combine them into a quasi-F ratio (minF') that afforded a test of the experimental effect incorporating both by-participants and by-items error terms. Analyses have shown that minF' analyses perform well in the sense that Type I errors are committed at a rate corresponding to the nominal .05 or .01 significance threshold (Baayen et al., 2008; Barr, Levy, Scheepers, & Tily, 2013). However, such analyses suffer from two critical limitations. Firstly, use of the approach is restricted to situations where data have been collected in a balanced fashion across the cells of the experimental design. Most researchers know that balanced data collection is rare. Experimenters can make mistakes and observations are missed or lost. Participants make errors and null responses may be recorded. Perhaps critically, in practice, Raaijmakers et al. (1999; Raaijmakers, 2003) showed how the use of minF' declined and was replaced by the reporting of separate F1 and F2 analyses, despite the associated risk of elevated Type I error rates (see also Baayen et al., 2008).

The minF', F1 and F2 analyses are also restricted to situations where data have been collected according to a factorial design. That is, comparing outcomes recorded for different levels of a categorical factor or different conditions of an experimental manipulation. However, researchers often seek to examine the relationships between continuous outcome and continuous experimental variables. Cohen (1983) demonstrated that the cost of dichotomizing continuous variables is to substantially reduce the sensitivity of analyses. This may be especially important

where the relationship between outcome and experimental variables cannot be assumed to take a monotonic function (Cohen, Cohen, Aiken, & West, 2003). In such circumstances, researchers have tended to estimate the effects of continuous experimental variables using multiple regression, e.g., predicting by-item mean reading response latencies from a set of predictors capturing different word properties (Balota et al., 2004). However, Lorch and Myers (1990) demonstrated that such by-items regression analyses reverse the language-as-fixed-effect problem by failing to take into account random between-participants differences.

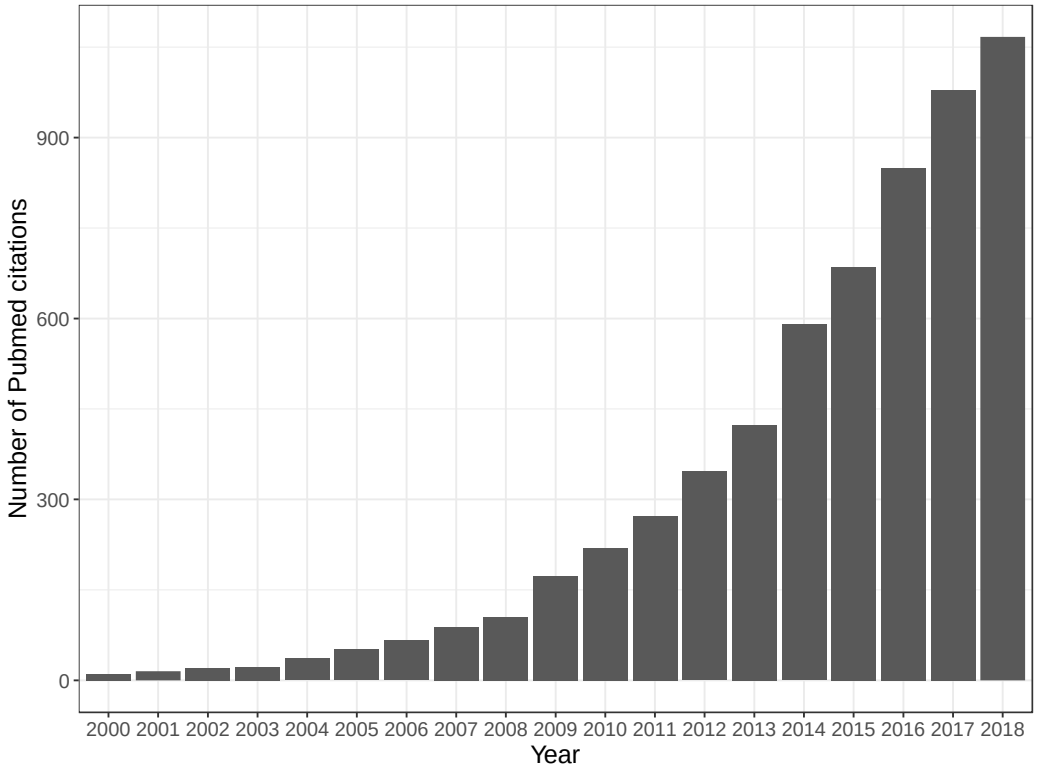
Lorch and Myers (1990) recommended that the researcher conduct a two-step analysis, firstly, conducting a regression analysis separately for each participant, e.g., predicting a participant's response latencies from variables indexing stimulus properties and then, secondly, conducting an analysis of the per-participant coefficients estimates. This approach, sometimes known as slopes-as-outcomes analysis, has been used in some highly cited experimental Psychology studies (see examples by Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Kliegl, Nuthman, & Engbert, 2006; Zwaan, Magliano, & Graesser, 1995) though perhaps more often in educational and other areas of social science research (see, e.g., citations of Burstein, Miller, & Linn, 1981; see discussion in Kreft & de Leeuw, 1998). However, the approach does not take into account variation between participants in the uncertainty about coefficients estimates (e.g., if one participant has fewer observations than another). That is, in a two-step analysis it is not possible to distinguish variation between per-participant coefficients and error variance (Snijders & Bosker, 2011). As well as avoiding the language-as-a-fixed-effect-fallacy, LMMs are also a solution to the limitations of slopes-as-outcomes analyses, as they 'shrink' – or pool - estimates towards sampling unit means (e.g., participant means) when there are fewer data points for that grouping (e.g., more missing data points for a participant, Gelman & Hill, 2007).

Introductions to LMMs (e.g., Snijders & Bosker, 2011) often discuss random differences between sampling units (e.g., between participants) either as error variance that must be controlled, or as phenomena of scientific interest (e.g., Baayen et al., 2008; Bolker et al., 2009; Gelman & Hill, 2007; Kliegl et al., 2011). Either way, LMMs allow this variation to be modeled by the experimenter as random effects. This means specifying that the measured outcome deviates, per sampling unit, from the average of the data set (random intercepts, see Figure 1) or from the average slope of the experimental or covariate effect of interest (random slopes, Figures 2-4). Random intercepts and random slopes variance estimates can tell us how much of the overall error variance is accounted for by variation between sampling units, e.g., the *differences* in overall RT between participants or between items. They can also tell us what the estimated difference is for a given sampling unit, e.g., by how much does a participant's overall RT differ from the grand mean RT?

It is worth highlighting that if these systematic differences in hierarchically structured data-sets are not properly accounted for, then false positive results become worryingly high (e.g., a Type I error rate as high as 80%: Aarts et al., 2014; see also Clark 1973; Rietveld & van Hout, 2007) and the power of summary statistics to detect experimental effects is reduced (Aarts et al., 2014). More generally, an analysis that fails to account for potential differences between sampling units in the slopes of experimental variables can mis-estimate the robustness of observed effects (Gelman, 2014). For example, one half of participants may show an effect in a positive direction and half show an effect in a negative direction. If this variation is not captured, the estimated direction of the average effect across all participants can be misleading (for an excellent exploration of this, see Jaeger, Graff, Croft, & Pontillo, 2011). Given these numerous analytic advantages, LMMs have been rapidly adopted, particularly in subject areas such as psycholinguistics (Baayen, 2008; Baayen et al., 2008).

----- FIGURE 5 Here --------

Figure 5 Title: Number of Pubmed citations for 'Linear Mixed Models' by year



1.3 So what is the problem?

The problem for researchers is that there are multiple analytic decisions to be made when using LMMs. This issue is not new to their advent in experimental Psychology. Simmons, Nelson, and Simonsohn (2011) demonstrated the decisive impact on results of 'researcher degrees of freedom'. Silberzahn and Uhlmann (2015) showed that the same data can reasonably be analysed in a variety of different ways by different research groups. Neither demonstration depended on the use of LMMs. The proliferation of alternate findings that arise from variation in choices at each point in a sequence of analytic decisions is crystallized by Gelman and Loken (2013) in the metaphor 'the garden of forking paths'. Multiple analytic steps make variation in observed results more likely, even when reasonable

assumptions and decisions have been made at each step (Gelman & Loken, 2013; Silberzahn & Uhlmann, 2015). The same concerns have arisen in fields other than experimental Psychology, for example, following the rapid expansion in neuroimaging studies in which complex analyses with multiple analytic steps are the norm (Carp, 2012a; Carp, 2012b; Poldrack & Gorgolewski, 2014; Wager, Lindquist, & Kaplan, 2007). Thus, this paper reports on the use of LMMs in the context of ongoing concerns regarding statistical best practices across the cognitive and neurosciences (e.g., Carp, 2012a, 2012b; Chabris et al., 2012; Cumming, 2013a, 2013b; Ioannidis, 2005; Kriegeskorte et al., 2009; Lieberman & Cunningham, 2009; Pashler & Wagenmakers, 2012; Simmons, Nelson & Simonsohn, 2011; Vul et al., 2009). As we shall report, the decisions that researchers must make when conducting LMMs appear to be associated with a heightened sense of uncertainty and insecurity.

Researchers' concerns may stem, in part, from the fact that the rapid adoption of LMMs has not been complemented by the adoption of common standards for how they are applied and, critically, how they are reported. There are many excellent introductory texts available for LMMs (e.g., Baayen et al., 2008; Baayen, 2008; Bates, 2007; Bolker et al., 2009; Bryk & Raudenbush, 1992; Gelman & Hill, 2007; Goldstein, 2011; Hox, 2010; Judd, Westfall, & Kenny, 2012; Kreft & de Leeuw, 1998; Pinheiro & Bates, 2000; Snijders & Bosker, 2011; see also Appendix 3). The caveat here is that even some texts that are designed to be introductory require a higher level of mathematical literacy than is required for or delivered by a majority of undergraduate psychology courses (e.g., fluency in linear and matrix algebra). It is also not clear how many undergraduate courses teach LMMs. Therefore, students may be required to read research papers that they are not equipped to understand. It may be feared that educational resources are sufficient to motivate the use of LMMs but are not sufficient to enable their appropriate application by researchers. Established researchers may baulk at the time needed to

undergo retraining in software applications and statistics, and to have to allocate more time in the future as software and analytic practices update.

The development of appropriate training for current or developing researchers is an important concern for the future but we are optimistic that this challenge can be met over time. There are a growing number of LMM tutorials available for different disciplines which include examples and technical descriptions of software use (Baayen et al., 2008; Brauer & Curtin, 2018; Brysbaert, 2007; Chang & Lane, 2016; Cunnings, 2012; Field, Miles & Field, 2009; Field & Wright, 2011; Jaeger, 2008; Kliegl, 2014; Magezi, 2015; Murayama, Sakaki, Yan, & Smith, 2014; Rasbah et al, 2000; Rabe-Hesketh & Skrondal, 2012; Schluter, 2015; Th. Gries, 2015; Tremblay & Newman, 2015; West & Galecki, 2011; Winter, 2013). From the authors' own experiences, as interested but not mathematically expert readers, the most friendly and relevant tutorials for language researchers can be found in Brysbaert (2007), Cunnings (2012) and Winter (2013). Once the reader is comfortable, we strongly recommend the recent paper by Brauer and Curtin (2018).

The trouble is not that researchers are not doing what experts advise but, rather, it is the ways in which researchers have responded to the evolution of recommendations in what is, in part, a methodological field with active areas of development. Critically, the literature on LMMs is fairly consistent in terms of recommendations for best practice but there has been some diversity in the guidance available to researchers (e.g., compare Barr et al., 2013; Bates, Kliegl, Vasishth & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Thus, a critique by Barr and colleagues (2013) of the application of relatively simple mixed-effects models, including random intercepts but not random slopes, led to a wider sense of unease about the replicability of previously published results. It appeared to many that the frequent reporting of findings from LMMs including just random intercepts would be associated with an inflated risk of false positives. However, latterly, it has been argued that the risk of false positives must be balanced with the risk of false

negatives through the inclusion of a parsimonious selection of random effects (Matuschek et al., 2017). This apparent diversity in recommendations could be a source of the uncertainty in approach or diversity in practice that (to anticipate) our observations uncover. But we would read the succession of publications as marking a progression in our understanding of the most useful application of mixed-effects models. Many agree that LMMs are appropriate to many experimental data analysis problems. Many assume that random effects should be incorporated in the LMMs that are fitted. As we will argue, the key issue is that since LMMs are an explicit modelling approach, they require a different attitude than has been ingrained, perhaps, through the long tradition of the application of ANOVA to the analysis of data from factorial design studies.

What we hope to make clear is that there is no single correct way in which LMM analyses should be conducted, and this has important implications for how the reporting of LMMs should be approached. Researchers will, quite reasonably, be guided in their approach to analyses by the research question, the structure of the data as it arises from study design and the process of observation, and the constraints associated with the use of statistical computing software. The problem is that variation in practice – especially reporting practice - can have a direct and damaging impact on our ability to aggregate data and to accumulate knowledge. Replicability and reproducibility are critical for scientific progress, so the way in which researchers have implemented LMM analysis must be entirely transparent. We also hope that the sharing of analysis code and data becomes widespread, enabling the periodic re-analysis of raw data over multiple experiments as studies accumulate over time.

1.4 Present study

We examine the diversity in practices adopted by different researchers when reporting LMMs, and the uncertainty that that diversity appears to engender. We completed a survey of current LMM practice in psychology. This consisted of two parts, a questionnaire sent out to researchers and a review of papers that used LMM analyses. We found widespread concern and uncertainty about the implementation of LMMs alongside a range of reporting practices that frequently omitted key information. The survey demonstrates the assimilation of a data analysis method in our discipline in 'real time'. To address these concerns, we present a set of best practice guidance with a focus on clear and unambiguous reporting of mixed-effects model analysis.

2.0 Questionnaire

2.1 Method

2.1.1. Participants

163 individuals completed the questionnaire: 94 females, 63 males and 6 who did not disclose their gender. Mean age was 36 years (standard deviation, SD = 9.26, range = 23-72). Just under 40% of respondents reported their discipline as Psycholinguistics, 16% Linguistics, 11% Psychology, 5% Cognitive Science/Psychology, 4% Language Acquisition and 3% Neuroscience; 15% of individuals reported more than one discipline. A number of other disciplines were reported by individuals (e.g., Anthropology, Clinical Psychology, Reading). Mean number of years working in a given discipline was 10.38 (SD = 8.16, range = 0.5-30).

Data on academic position and institution can be found in Table 1. We recognize that this sample is biased towards those already using mixed-effects models (rather than reading about them), and this was reflected in the high proportion who stated they already used them (see 2.2.1 below) and the small number who stated they were planning to use mixed-effects models (2.2.3).

Table 1: Reported position and institution type for questionnaire respondents (% of total)

Position	**%**	**Institution**	**%**
Undergraduate	0.00	University UK	25.77
Postgraduate MSc	1.23	University Other	59.51
Postgraduate PhD	24.54	Research Institute UK	0.61
Postdoctoral researcher	26.38	Research Institute Other	9.82
Lecturer/Assistant Professor	24.54	Institution Other	4.29
Reader/Senior Lecturer /Associate Professor	11.04		
Professor	9.20		
Other	3.07		

2.1.2 Design and procedure

A qualitative questionnaire was used, with both open and closed questions. Ethical approval for the study was granted by the University of Reading School of Psychology & Clinical Language Sciences Research Ethics Committee. The online questionnaire was implemented in LimeSurvey (LimeSurvey Project Team & Schmitz, 2015). Individuals were invited to complete the questionnaire via email lists and personal emails to academic contacts of the authors, with a request to forward on to any interested parties. All responses to the questionnaire were anonymous. The questionnaire began with a brief introduction to the study. Consent was provided by checking a box to indicate agreement with a statement of informed consent.

The questionnaire elicited answers to questions focusing on the use and reporting of Linear Mixed-effects Models (LMMs). Appendix 1 provides the full questionnaire. Questions covered demographic information, use and reporting of

LMMs, challenges encountered when using LMMS and concerns they had for their own research and their field.

A period of approximately one month was allowed for responses to be collected. Data collection was stopped once we had reached the current sample size, as the sign-up rate to complete the questionnaire had slowed. The sample size was judged adequate for our purposes (frequency and thematic analysis of question responses) and we judged that a substantial increase in numbers was unlikely if we left more time.

2.1.3 Analysis

The complete data can be found at osf.io/bfq39 – Files – Mixed models survey results_analysis.xlsx. For closed questions, the percentage of responses falling into a given category were calculated. For open-ended questions, thematic analysis was completed to identify the most common responses across individuals (Braun & Clarke, 2006). Individual responses to each question (e.g., challenges to using LMMs) were collated as rows in a spreadsheet and given a thematic label to code the response (e.g., software, convergence, lack of standard procedures etc.). Responses were then reviewed and sorted, combining responses that fell into the same thematic label. We were interested in reporting the most common responses, so the total number of responses that fell into a given theme were counted as a percentage of the total responses to that question. For questions where categorical responses were made (e.g., reporting software used, listing training and resources), we generated lists of unique responses and the frequency (% of total) with which each one was reported. The results of the above analyses are presented together.

2.2 Results

2.2.1 Usage of mixed-effects models

The great majority of respondents (91%) had used mixed-effects models for data analysis. The mean year of first using LMMs was 2010 (SD = 3.94 years, range = 1980 – 2014). We asked respondents to estimate how often they used mixed-effects models, the mean was 64% of data analyzed (SD = 31%, range = 0-100).

2.2.2 Training & software

The majority of respondents had attended a workshop, course or training event to learn mixed-effects models (68%), 30% had learnt from colleagues, 21% from internet resources, 12% using specific books or papers, 10% were self-taught and 9% had learnt from a statistics advisor or mentor. Appendix 2 provides a comprehensive list of the specific authors, papers, books, websites and other resources used by respondents. Readers may find this useful for their own training needs.

The majority of individuals used the statistical programming language and environment, R (90%) (R Core Team, 2017), with 20% mentioning the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Other named R packages were gamm4 (Wood & Scheipl, 2016), languageR (Baayen, 2013), lmerTest (Kuznetsova, Brockhoff & Christensen, 2016), mgcv (e.g., Wood, 2011), and nlme (Pinheiro et al., 2016). The next most frequently used software was SPSS (8%; IBM Corp, 2013). A number of other software applications were named by one or two people: MLwiN (Rasbash et al, 2009), Matlab (Matlab, 2012), Mplus (Muthén & Muthén, 2011), Stan (Stan Development Team, 2016), JASP (JASP Team, 2016), S-PLUS (e.g. Venables, 2014), SAS and ESS (Rossini et al., 2004).

2.2.3 Planned use

For individuals who had not yet used mixed-effects models (15 respondents),

11 were planning to use them and four were not. For those planning to use them, reasons included exploration of a larger number of predictor variables (5 responses), to look at change over time or longitudinal data (2 responses) and for better statistical practice (e.g., control of individual differences, inclusion of random effects; 2 responses).

2.2.4 Challenges to using mixed-effects models

The most frequently reported concern was a lack of consensus or established, standardized procedures (26% of respondents; e.g. "it's quite difficult… to understand what standard practice is"). Related to this were responses that described a lack of training or clear guidelines for analysis, interpretation or reporting results (13%; e.g., "minimal training/knowledge available in my lab", "Presentation of data for publication") and the relative novelty of the analysis (7%; "it is relatively new so recommended practices are in development and not always fully agreed upon"). A number of responses highlighted a lack of knowledge (18%; e.g., "I do not know enough about them", "some reviewers request these models but researchers are not all skilled in these techniques", "complex math behind it not easy to grasp", "not enough people who know it"). A broad challenge in applying conceptual knowledge was seen in responses covering difficulties in selecting or specifying models (25%; e.g., "Model specification - knowing what to include and what not to include"), models which fail to converge or in which assumptions are violated (14%; e.g., "How to deal with models that fail to converge"), understanding and interpreting random effects structures (16%; e.g., "Determining what constitutes an appropriate slope term in the random effects structure"), identifying interactions (4%; "Working out significant interactions") or interpreting results generally (7%; "difficulty in interpreting the results"). Other specific points included models being overly flexible or complex (e.g., "The potential complexity of the models that goes substantially beyond standard procedures", "Mixed models are so flexible that it can be difficult to establish what is

the best suited model for a given analysis") and challenges in checking and communicating model reliability (e.g., "Knowing how to test whether a model violates the assumptions of the specific model type being used"). The most frequently reported concerns are reported in Table 2.

Table 2: Most frequently reported challenges and concerns in using LMMs*

Reported challenge	**%**
Lack of standardized procedures	26
Selecting and specifying models	25
Researcher reports lack of knowledge	18
Understanding and interpreting random effects	14
Lack of training/guidelines for analysis, interpretation and reporting	13
Use of new and unfamiliar software	12
General concern over use of LMM for own analysis	75
Reporting results	15
Model selection	14
Learning and understanding analysis	14
Lack of established standards	11
General concern over use of LMMs for discipline	73
Lack of standards	23
LMMs used when not fully understood	23
Misuse of models	17
Reporting is inconsistent and lacks detail	17
Peer review of LMMs is not robust	10

*identified by thematic analysis

Technical challenges were highlighted, specifically the use of new or unfamiliar software (12%; e.g., "software package (R) I was unfamiliar with") and the

reliability of analysis code (e.g., "Some of the code might also not be reliable. For example, people reported differences when running the same analysis in different versions of the same software"). A number of individuals reported specific difficulties with model coding and fitting (e.g., coding of categorical variables, setting up contrasts, structuring data appropriately, forward and backward model fitting and post-hoc analyses).

A number of responses reflected unease at the shift from traditional factorial designs and ANOVA or other inferential statistical tests (e.g., "[lack of] convincing evidence that mixed models provide information above and beyond F1 and F2 tests"). For example, susceptibility to p-value manipulation or difficulties in establishing p-values (4%; “too many people still believe that we are fishing for p-values if we do not use classical anovas"), knowing how to map models onto study design (4%; "Knowing when it's appropriate to use them", "to understand the influence on future study designs"), difficulties with small samples, sparse data and calculating effect sizes or power.

2.2.5 Concerns using mixed-effects models for own data and in the wider discipline

Around 75% of respondents had concerns over using LMMs in their own data analysis. For these respondents, the most salient concerns were reporting results (15%; e.g., "Do you report your model selection criteria and if so, in what level of detail… perhaps several models fail to converge before you arrive at one that does?"), selecting the right model (14%; e.g., "model selection"), learning how to do the analysis and fully understanding it (14%; e.g., "I do not have enough knowledge to correctly apply the technique"), a lack of established standards (11%; e.g., “the lack of standardized methods is a problem"), models that do not converge (9%; e.g., "How to deal with convergence issues") and the review process when submitting LMM analysis for publication (9%; e.g., “experimental psychology reviewers are often

suspicious of them"). Other concerns broadly reflected those already identified as challenges above. See Table 2 for the most frequently reported concerns.

Around 73% of respondents had concerns over the use of LMMs in their discipline or field. Here, the key concerns were a lack of standards (23%, e.g., "lack of established standards"), use of models without them being fully understood (23%; e.g., "Overzealous use of random effects without thinking about what they mean"), frank misuse of models (17%; e.g., "Misapplication of mixed models by those not at the forefront of this area"), reports of model fitting being inconsistent and not detailed enough (17%; e.g., "not describing the analysis in enough detail"), a lack of familiarity and understanding of the models (10%; e.g., "lack of knowledge about their implementation") and the review process not being robust (10%; e.g., "Reviewers often can't evaluate the analyses"). Additional concerns were over researchers being able to misuse the flexibility of mixed-effects models (5%; e.g., "increased 'researcher degrees of freedom' ") or "p-hack" the data (3%; e.g., "It's easier to p-hack than an ANOVA"), and the breadth of approaches to making decisions during model fitting (4%; e.g., "The variety of approaches people take for deciding on model structure"). There was also concern over why LMMs were deemed better than factorial ANOVA approaches (3%; e.g., "Why are they privileged over simpler methods?") and that it was difficult to compare them against these traditional approaches (2%; e.g. "less accessible to readers/reviewers without experience… than traditional analyses"). See Table 2.

2.2.6 Current practice

For respondents who were currently using mixed-effects models, 70% did not specify variance-covariance structures for the models. On reflection, participants may not have understood this question given that it was not accompanied by an explanation of these terms. We asked people to provide a typical model formula from their analyses. Two individuals stated that they used SPSS, and therefore did not

specify model formulae. Of those who did provide an example, only three explicitly mentioned model comparison and model checking. See Table 5 for a summary of random effects from model examples. 100% specified random intercepts for subjects/participants and 92% specified random intercepts for items/stimulus materials or trials. Random slopes to allow the effect of interest to vary across subjects and/or items were less common (62%).

Table 3: Current practice

Current practice	**%**
Do you specify variance-covariance structures?	
Yes	30
No	70
Random Effect structures from model examples:	
Random intercepts for subjects	100
Random intercepts for stimuli/trials	92
Random slopes for effect to vary across subjects	62
Comparison to factorial analysis (ANOVA)	
Do you compare LMMs to factorial analysis?	
Yes	61
No	24
N/A	15
Were results comparable?	
Yes	33
No	46
N/A	21
How do LMMs compare to factorial analysis?*	
LMM are better fit to data	28
Largely comparable	26
LMMs are more sensitive/less conservative	16
LMMs are more conservative	8

*identified by thematic analysis

When included, random slopes were often qualified on the basis of experimental design and only included when appropriate for the data structure (e.g., random slopes for within-subject factors; Barr et al., 2013). Where multiple predictor factors were included, interactions between factors for random slopes were typically included. It is notable that some respondents stated that they did not include interaction terms for random slopes, excluded these first if the model failed to converge, or simplified random effects until the model converged. Some removed the modeling of correlations between random effects for the same reason. See Table 3.

2.2.7 Comparison to traditional approaches

Around 61% of respondents had compared the results of LMM analyses to the results of more traditional analyses (i.e. ANOVA or other factorial inferential statistics; 15% responded N/A). Of those, 33% reported that results had been comparable, 46% reported that results were not comparable and 21% responded with N/A. An open question asked for respondents' evaluation of this comparison. The most frequent response was that results were comparable (26%; e.g., "Largely methods correspond to each other"). A number of responses identified that mixed-effects models were preferred or gave a better, more detailed fit to the data (28%; e.g., "I think we got a better fit for our data using LMEs instead of the traditional ANOVAs/Regression models"). However, it is not clear whether results were comparable in terms of the size of numeric effects or coefficients. Responses instead focused on whether results were significant. LMMs were reported to be more sensitive/less conservative, demonstrating significance for small effects (16%; e.g., "differences can occur if effects are just above or below p=.05", "mixed models seems less conservative than for example (repeated measures) anova"). However, mixed-effects models were also found to be more conservative, depending on how

the random effects structures were specified (8%; e.g., "Mixed models are typically more conservative, but not always"). Traditional F1/F2 tests were sometimes used to confirm or interpret effects in the mixed-effects models (4%; "I look if both analyses point to the same effects of the experimental manipulations") and in one instance F1/F2 tests were reported to be “much easier and less time-consuming” than LMMs. See Table 3.

2.2.8 Reporting & Preferred reporting

Respondents were asked for their typical practice when reporting models, this question was multiple choice and a summary of responses is given in Table 4. The vast majority reported p-values and model fitting (88% and 80% respectively), but other options were chosen much less often: model likelihood was reported by 50% of respondents; confidence intervals by 37%; specification/reporting of model iterations by 36%; and F-tests between models by 31%.

Table 4: Current practice in reporting mixed models (% total)*

What is reported	% Yes	% No
p-values	88	12
Model fitting	80	20
Likelihood	49	51
Confidence intervals	37	63
Iterated models	36	64
F-tests	31	69

*ordered by frequency of response high to low, rounded up to nearest %; 147 responses. Respondents were asked simply to indicate whether they reported model fitting and model likelihood, for detailed discussion of these parts of LMM analysis see sections 4.1.4 to 4.1.6.

For preferred reporting format, the majority were in favour of a table (53%), followed by written information in the text (19%) and then plots (15%). The main reasons for selecting tables were ease of reading and clarity. Written text could provide details and facilitate interpretation. Plots were deemed important for more

complex models and to visualize the model structure. Some individuals stated that reporting format should depend on the data and model complexity (7%).

2.2.9 Sharing of Code and Data

We asked respondents to state whether they would share data and code, with 70% responding that they would share both (e.g., “Yes. Science should be open in its practice”). Table 5 summarizes the responses. Some respondents specified that they would share data only after publication, on request, after an embargo or when they had finished using it (9%; e.g., “I would be willing to share data on personal request”). Reasons for sharing included being open and transparent or a duty to share work that had been publicly funded (e.g., “yes, always. No-brainer: tax-payer-funded scientist”). A number of respondents identified a general benefit to the field and to improve standards. For example, to contribute to meta-analysis or further data exploration (e.g., “… to facilitate additional research and replication of previous results. This data would also be extremely helpful for meta-analyses and for future research to be able run power analyses based on previous findings”). Analytic rigour was also mentioned, for example having a more open discussion about how models are used, checking model fitting, correcting errors, and having more experienced people look at the data (e.g., “We definitely need transparency and standards here because most of us are not statisticians”). Around 3% would not share data and 3% were unsure. Reasons included not wanting to be ‘scooped’, and being unsure if data sharing was allowed on ethical grounds. One respondent asked “Why should I share my data?”.

Table 5: Sharing of code and data*

Would you share data and code?	**%**
Share both data and code	70
Share code	15
Specified sharing of data after publication	9
Would not share either	3
Unsure	3
Would you like access to data and code?	%
Access to both	74
Access to both but unlikely to use it	6
Access to code	9
Did not want access to either	3
Did not want access to code	3
Did not want access to data	2
Unsure	2

*Identified from thematic analysis

Around 15% responded that they would share code, with no statement about data sharing. Reasons for sharing code included it being good practice and good for learning, as well as comparing analyses (e.g., “Good practice, other researchers can look at what you did and learn something, or point out errors”, “I think it is helpful to share code. This will hopefully lead to a more open discussion of the choices we all make when doing this type of analyses”). Two individuals stated that they would not share code due to their inexperience. A few respondents mentioned difficulty in sharing code that could often be ‘messy’ and that it would be time consuming to prepare code for publication.

We asked respondents to state if they would like to access data and code in published reports. Around 74% would like access to both, with a further 6% specifying yes but that they would be unlikely to use it. Reasons for accessing were broadly similar to those identified above, with mention of transparency, improved

standards, for learning, for meta-analysis, analytic rigour and checking reported data. Some respondents reported that current data sharing practices were already sufficient (e.g., sharing data on request, depositing in centralised archives, e.g., “Doesn't have to be in published reports. Can be in a database accessed via the publisher or institute”), or that this was a wider issue and not specific to LMMs (e.g., “I don't see the access to data and code being a mixed effects specific issue. This is for any paper, regardless of the statistical technique used”). A smaller number specified that they would like access to code (9%) with no statement about data. Some respondents qualified that data and code should be part of supplementary materials or a linked document, rather than in the publication itself. Finally, a few people did not want access to code (3%), data (2%) or both (3%), or they were unsure (2%). See Table 5.

2.3 Discussion

Most respondents had concerns over the use of LMMs in their own analyses and in their discipline more widely. Concerns were driven by the perceived complexity of LMMs, with responses detailing a lack of knowledge (own knowledge, that of reviewers or other researchers). Our interpretation is that this knowledge deficit (perceived or real) drives the other concerns. Namely, difficulties in learning and understanding the analysis process and difficulties in building, selecting and interpreting LMMs. For some, these difficulties are compounded by having to learn about new software applications (for an overview of software applications and their comparability see McCoach et al., 2018). Software applications undergo changes and updates which may change the results of a fitted model, as illustrated in the grey literature around lme4 (e.g. internet discussion boards such as stackoverflow.com; Nava & Marius, 2017). Such back-end changes – typically not salient to the average

psychology researcher – will add to the sense that mixed-effects models are complex and problematic. Respondents were concerned by not knowing what to report or how to report results from LMMs. This point feeds into reports of LMMs being received skeptically by reviewers as inconsistent formatting and presentation of analyses will exacerbate difficulties in the review process. It may be that two individuals trained in LMMs complete analyses that are true to their original training, but which – for similar data – differ in implementation and are reported differently in publications. Given that reviewers are sampled from the community of active researchers, lack of knowledge in reviewers was also a concern. At present, we are using a method of analysis that the community feels is not well understood, not clearly reported and not robustly reviewed. Little wonder that some see it as overly flexible and yet another way of fishing for results.

Most researchers report p-values for model coefficients and some detail of model fitting for LMMs, fewer provide details of iterated models or Likelihood comparisons between different models. This means that, in general, the number of decisions being made during model fitting and the process of model selection is not transparently reported in manuscripts. This lack of transparency should not be seen as deliberate obfuscation: most respondents were willing to share analysis code and data, and felt that it was important to do so.

Alternate choices taken at multiple analytic steps can foster the emergence of different results for the same data (Gelman & Loken, 2013; Silberzahn & Uhlmann, 2015) giving the impression of unprincipled flexibility. The rapid uptake of LMMs has been driven, in part, by the need to explicitly account for both subject- and item-related random error variance (Locker, Hoffman & Boviard, 2007; Baayen et al., 2008; Brysbaert, 2007) and part of the anxiety over model building arises when one moves from factorial ANOVA into LMMs (Boisgontier & Cheval, 2016). Although ANOVA and LMM share a common origin in the general linear model, they are very different in terms of execution. In LMMs, the analysis process is similar to regression

(Bickel, 2007). A model equation for the data is specified and reliable analysis requires larger data sets (e.g., trial-level data or large samples of individuals, Baayen, 2008; Luke, 2016; Maas & Hox, 2004; 2005; Pinheiro & Bates, 2000; Westfall, Kenny & Judd, 2014). Nested models may be compared or 'built' to find the best fit to the data. The process feels notably different to producing a set of summary statistics (e.g., averaging responses to all items for a subject), which are then put through a factorial analysis (such as ANOVA). Survey responses reflected this uneasy shift. Of respondents who had compared LMMs to ANOVA, a third found comparable results but nearly half found results that were not comparable. For those who had compared the two analyses, LMMs were reported to be a better fit to the data, but could be both more or less conservative especially when effects were marginally significant under ANOVA. It is worth noting that LMMs are not a new level of complexity for statistics in cognitive science (compare: structural equation modelling, Bowen & Guo, 2011; growth curve modelling, Nagin & Odgers, 2010), especially when compared against advances in brain imaging analysis and computational modelling. However, the perceived complexity is demonstrated by survey responses repeatedly referring to a lack of knowledge and established standards.

The survey data clearly demonstrates that researchers are uncomfortable with the use of LMMs. This is despite a number of excellent texts (see Appendix 2, and references given in the Introduction) and an explosion of online tutorials and support materials. To evaluate whether there is a problem in how LMMs are actually implemented and communicated, we completed a review of published papers using LMM analysis.

3.0 Review of current practice in use and reporting of LMMs

Our objective was to review current practice in the use and reporting of LMM/GLMMs in linguistics, psychology, cognitive science and neuroscience. This complements the survey by adding objective data on how LMMs are used and reported.

3.1 Method

We completed a review of published papers using LMM analysis, taking a sample rather than exhaustively searching all papers. This approach was chosen to make the review manageable. To start, the first author used Google Scholar to find papers citing Baayen et al. (2008), widely seen as a seminal article whose publication was instrumental to the increased uptake of LMM analysis (cited over 3500 times to date). To keep the review contemporary, papers were chosen from a four-year period spanning 2013, 2014, 2015 and 2016. Papers had to be in the field of language research, psychology or neuroscience (judged on the basis of title, topic and journal). From each year, the first 100 citations fitting the above criteria were extracted from Google Scholar, when limited by year, giving 400 papers in total. The first search was completed on 30th May 2017, giving a total of 3524 citations for Baayen et al (2008) with 2360 citations between 2013-2016. Therefore, we sample ~17% of the papers fitting our criteria, published in that four-year period. Sixteen papers were excluded as they did not contain an LMM analysis (e.g., citing Baayen et al. in the context of a review, or when referring to possible methods). One paper was not accessible. Three papers were initially reviewed to establish the criteria for classifying papers, with an excel spreadsheet created with a series of drop-down menus for classification. To check coding and classifications, the second author looked at one reported model from 80 papers (20% of the total papers coded; 20 papers from each year). Initial agreement was 77%, with differences resolved by

discussion. The spreadsheet with all the data and classifications from the review can be found here (https://osf.io/bfq39/; Files – Baayen Papers Rev with coding check.xlsx). Classification criteria are summarized in Table 6, and a fuller description of these can be found in Appendix 3.

Table 6: Classification criteria for review and associated data table

Criteria	Options	Data Table
Field / Topic	Psychology, Linguistics & Phonetics, Neuroscience, Psycholinguistics.	
Model Type	LMM, GLMM, LMM & GLMM, GAMM, Other.	A4.1
Approach	ANOVA testing for fixed effects via LRTs/model comparison ANOVA testing with random effects of interest Regression with random effects control for subject / item variance Regression with multiple predictors and control variables Regression with random effects of interest Repeated measures / control for hierarchical sampling Repeated measures with random effects of interest	A4.2
Model Comparison	LRTs, AIC/BIC, LRTs & AIC/BIC, descriptive	A4.3
Statement on model selection	What detail is given by the authors on how different models have been compared, or a final model selected?	A4.4
Convergance / Random Effect simplification	What detail is given by the authors of any convergence issues and what was done to deal with this (e.g. model simplification)?	
Model equation	Yes reported, not reported, given for some and not others	A4.5
Dependent variable	RT, Errors / Categorical variable, RT & Errors, eye movement data, brain imaging data, other	
Fixed Effects 1	IV, IV & Control variables	A4.6
Fixed Effects 2	Main effects, main effects & interactions	
Random Effect approach (if mentioned)	LRTs; LRT & AIC/BIC; LRTs/AIC for slopes; Maximal structure; LRTs backwards from maximal, LRTs upwards from minimal; LRTs against null	A4.7
Random Effect Intercepts modelled	Subject, Item/other, Subject & Item/other, Subject, item & other, Item & other	
Random Effect Slopes modelled	FE over subject, FE over item/other, FE over subject & items/other, FE over subject with interactions, FE over items/other with interactions, FE over subject &	

	items/other with interactions	
Random Effect covariances modelled	Yes reported as modelled, no not modelled, unclear whether modelled or not	
Reporting Format	Text only Text & Tables Text, tables & figures Table & Figures Text & Figures Figures Tables	A4.8
Reporting Fixed Effects	Coefficients Coefficients, t & p Coefficients, SE/CI Coefficients, SE/CI, t/z Coefficients, SE/CI, t/z & p Coefficients, SE/CI, p Coefficients, p t/z, p p *Additional note if condition means reported, not coefficients*.	A4.9
Reporting Random Effects	Variance, variance & covariance, or not reported	A4.10
Model fit reported	R2, model estimate correlation with data, R2 & est. correlations, AIC/BIC, Log Likelihood, other (define), no.	A4.11
P-values (if mentioned)	Assume t > 1.96 / 2 MCMC LRTs F tests Sattherwaite Kenward-Rogers	A4.12
Appendices for full reporting (if mentioned)	Yes	A4.13

3.3 Results

The complete data set can be found at osf.io/bfq39 and tables with counts in Appendix 4. Here we will summarize the data by walking through the stages of LMM analysis: model selection, evaluating significance and reporting results. Tables presenting counts in Appendix 4 follow the order below.

3.3.1 Model Selection

The majority of papers used LMM (n=193), GLMM (n=88) or a combination of both LMM and GLMM (n=95). General Additive Models (GAMs) were rare in our sampled papers (n=5; see Table A4.1 in Appendix 4).

The majority of papers approached the use of LMMs as a variant on regression with random effects controlling for participant and item variation (n=272) but a number also used LMMs as a replacement for ANOVA (n=61). It was relatively rare for studies to look at the random effects as data of interest (n=13; see Table A4.2). The classic use of LMMs for hierarchical sampling designs was present relatively infrequently (n=26), which may be a result of the sampling process. LMMs have been used for a number of years in educational and organisational research to address questions concerning hierarchical sampling designs (Gelman & Hill, 2007; Scherbaum & Ferreter, 2009; Snijders, 2005). Baayen et al. (2008) – our seed paper - presents LMMs as a method to control for by participant and by item variation in experimental cognitive science.

Reporting the model selection process was infrequent (typically present in ~20-25 papers in each year; Table A4.3) and a wide variety of practices were present. Manuscripts reported “best fit” models following Likelihood Ratio Tests (LRTs) or Akaike Information Criterion or Bayesian Information Criterion (AIC/BIC) comparisons (n=23) or minimal model approaches in which models were simplified by removing fixed or random effects that were not significant (n=31). Models were also selected by moving from maximal to minimal models (n=6) or minimal to maximal models (n=8), or using backwards fitting (n=7).

Model comparisons for fixed effects were not present in all manuscripts (typically present in ~50-60 papers in each year; Table A4.4). This may be because researchers using experimental designs are modelling all fixed effects together (as in an ANOVA) rather than using model comparison to select them. When comparisons were present, the majority reported LRTs (n=129), with fewer reporting AIC or BIC

(n=12) or a combination of LRTs and AIC/BIC (n=20). Some papers described the model comparison process but did not provide data for the comparisons (n=54).

Model comparisons for random effects were also not present in all manuscripts (Table A4.5). The numbers that did test for the inclusion of random effects increased over time (2013 = 16, 2014 = 33, 2015 = 43, 2016 = 42). When a specific approach was reported, there was a clear preference for using a maximal random effects structure (Barr et al., 2013; n=86), followed in frequency by a preference for using Likelihood Ratio Tests to determine random effects structures (LRTs, n=25). Less common was a combination of starting with a maximal structure and then using LRTs to simplify (n=11) or starting with a minimal structure and using LRTs to add more complex random effects (n=7).

Reporting of convergence issues was increasingly common over the four-year period (2013 = 2, 2014 = 8, 2015 = 14, 2016 = 21; Table A4.4), and a variety of methods were reported for dealing with this. For example, simplification by removing slopes (n=9), correlations between intercepts and slopes (n=2) or both slopes and correlations (n=4). Some manuscripts reported the “fullest model that converged” without specific detail on how simplification took place (n=14).

Fixed effect predictors (Table A4.5) were most often modelled as main effects and interactions (n=287) as compared to main effects alone (n=94), the inclusion of control variables was also common (n=109). The vast majority of models included random intercepts for both participants and items (Table A4.6, n=277), with a good number that included intercepts for participants only (n=64). Random slopes were present in around half the papers (2013 = 41, 2014 = 50, 2015 = 67, 2016 = 58; Table A4.6). Most commonly, random slope terms were included to capture variation in fixed effect predictors varying as main effects over participants (n=78) or over both participants and items (n=94). It was less common to include the variation of fixed effect interactions as slopes over subjects and/or items (n=36). Where random slopes were modelled, it was rare for manuscripts to explicitly report whether

correlations or covariances between intercepts and slopes had been modelled (~10-15 papers per year) and this information was often unclear or difficult to judge (n=63).

A simple way to report the structure of a model is to provide the model equation (Table A4.7); this was given in a minority of papers with a clear increase over time (2013 = 7, 2014 = 6, 2015 = 26, 2016 = 22, total n = 61). However, the majority of papers did not provide this information (n=317).

3.3.2 Evaluating significance

We classified 10 different combinations or approaches to evaluating significance for fixed effects (see Table A4.8). It is worth noting that only around half the papers reported the method used (n=207), so we can assume that researchers employed methods that were defaults for software packages. The main methods reported were: MCMC bootstrapping procedures available in R (n=71); assuming t was normally distributed and taking t>1.96 or t>2 as significant (n=52); or taking p-values for fixed effects from Likelihood Ratio Tests (LRTs) comparing models with and without the effect of interest (n=40). Other options for evaluating significance involved using approximations for calculating degrees of freedom (e.g., Sattherwaite, n=20; number of observations – fixed effects n=2), or using F tests calculated over the model output (n=23) (see further discussion in Section 4.1.5 and 4.1.6).

It was very rare for measures of model fit to be reported (Table A4.9), with most papers not providing this information (n=330). When model fit information was provided, it was most often the Log Likelihood or AIC/BIC value (n=35) which are informative relative to another model of the same data. R^2 was provided in few cases (n=8).

3.3.3 Reporting results

Manuscripts typically used text, tables and figures to report model output (n=151) although other options were evenly split over text and tables (n=85) and text

and figures (n=94), with several only reporting model output in the text (n=52; Table A4.10). Thus, many papers do not provide a summary of model output in a table, as you would expect for an analysis using multiple regression.

We saw every possible variation in reporting fixed effects (Table A4.11). The majority reported fixed effect coefficients, standard errors or confidence intervals, test statistics (t/z) and p-values (n=128). It was also common to report the coefficients and the standard error or confidence intervals with a test statistic but no p-value (n=52), a p-value but no test statistic (n=39), coefficients without standard errors or confidence intervals (n=73), or to provide *only* a test statistic or a p-value (n=43).

Most studies did not report random effects at all (Table A4.12, n=304), with only 51 papers reporting variances and 23 reporting variances and correlations or covariances.

A small number of papers used appendices to provide a complete report on model selection, fitting and code used for analysis (n=25, Table A4.13).

3.4 Discussion

The variation in practice evident from the review of papers mirrors the uncertainty reported by surveyed researchers. Naturally, some of the variation will be attributable to what is appropriate to the data and the hypotheses (e.g., the use of LMMs or GLMMs, the modelling of main effects only or interactions). What concerns us is the evidence for unnecessary or arbitrary variation in the reporting of LMMs. Because it is arbitrary, this variation will make analyses difficult to parse and it will incubate an irreducible difficulty (given low rates of data or code sharing) for the aggregation or summary of psychological findings. This difficulty will, necessarily, impede the development of theoretical accounts or practical applications.

Prior to completing this work, we hypothesized that models were being used in different ways by the research community – as an alternative to multiple

regression or as an alternative to ANOVA. We found some support for this, the vast majority of models (70%) were framed as regression analyses, and around 15% as ANOVA analysis. We also found other approaches, for example, whether the random effects were reported as data of interest, or whether the study was explicitly controlling for a hierarchical sampling procedure. Around 56% of the papers reported some form of model comparison but did not always then give informative detail. For model selection, 24% provide explicit detail on the approach taken for fixed effects and around 35% provided detail on how the random effects structure had been chosen. The review of papers clearly shows both diversity of practice and a lack of transparency and detail in reporting. This makes the diversity confusing rather than a source of information. In this context, it is not surprising researchers report confusion and lack of knowledge.

Of particular interest was the variation in how significance was established. Only half the papers reported the method used, yet we encountered 10 different methods for testing significance in use. Depending on the study (e.g., confirmatory hypothesis testing or data exploration) researchers will have different needs for their analysis (Cummings, 2012). When replacing ANOVA or ANCOVA, researchers might want something similar to an F test that provides a p-value for the main effect or the interaction effect. This can be achieved by testing to see if the inclusion of a predictor improves model fit (e.g., Frisson et al., 2014; Trueswell et al., 2013). Alternatively, an ANOVA can be used to get F-tests for predictors. Here, the ANOVA summarises the variation across levels of a predictor in the model, and therefore how much variation in the outcome that predictor accounts for (e.g., if there is zero variation across experimental conditions, that manipulation does not change the outcome; Gelman & Hill, 2007). It is interesting to note that Gelman and Hill (2007) suggested the latter use of ANOVA not as a final analysis step in establishing significance, but as a tool for data exploration to inform which predictors are interesting when building models.

We found 63 papers that evaluated significance by using F tests or model comparison (~30% of the papers that reported a specific method of testing significance). However, it was not the case that LMMs framed as ANOVA always used this method for evaluating significance: such cases were evenly split across analyses framed as ANOVA (n=30) and those framed as regression (n=31, see Table A4.14). Where the analysis was framed as regression, we expected that it would draw on the power of LMMs to account for nested sampling groups (e.g., geographic or genealogical relationships between different languages, Jaeger, Graff, Croft, & Pontillo, 2011), modelling the influence of individual differences (e.g., such as age, Davies et al., 2017), change over time in repeated measures data (e.g., Walls & Schafer, 2006), or accounting for multiple predictor variables (Baayen & Milin, 2010; Davies et al., 2017). What researchers might want here is more similar to regression, exploring model building and comparison (e.g., Goldhammer et al., 2014) and coefficients for predictor variables. The vast majority of manuscripts were framed as regression and reported the significance of coefficients (n=122). Interestingly, it was almost never the case that papers reported both whether a coefficient was significant *and* whether the inclusion of that predictor improved model fit (n=2).

4.0 General Discussion

Linear Mixed-effects Models (LMMs) have, for good reason, become an increasingly popular method for analyzing data across many fields but our findings outline a problem that may have far-reaching consequences for psychological science even as the use of these models grows in prevalence. We present a snapshot of what psychological researchers think about mixed-effect models, and what they do when they publish reports based on their results. A survey of

researchers reveals that we are concerned about applying LMMs in our own analyses, and about the use of LMMs across the discipline. These widely-held concerns are linked to uncertainty about how to fit, understand and report the models. We may understand the reasons why we should use them but many among us are unclear how to proceed, as writers or as reviewers, in the absence of clear guidance, and in the face of marked inconsistencies in reporting practices. These concerns are mirrored in a striking diversity apparent in the ways in which researchers specify models, present effects estimates, and communicate the results of significance tests.

We observe that it is the reporting of models that is the principle point of failure. We find substantial, seemingly arbitrary, variation across studies in the information communicated about models and the estimates derived from them. We predict that this variation will make analyses difficult to parse, and thus will seed an irreducible difficulty for the future for the accumulation of psychological evidence. We saw that model equations were very rarely reported, though this is a simple means to communicate the precise structure of both fixed and random effects. Papers using LMM analysis do not always provide a complete summary of the model results. Fixed effect coefficients were not always reported with standard errors or confidence intervals. Random effects were hardly reported at all. These are all essential data for meta-analysis and power analysis. Curiously, then, the reporting of LMMs often ignores the key reason for using the analysis in the first place: an explicit accounting for the variance associated with groupings (sampling units) in the data. Random effect variances and covariances allow us to see just how much of the variance in the data can be attributed to, for example, individual variation (e.g., fast or slow participants) and the predicted effects (e.g., do fast participants always show a smaller effect?). If we care about psychological mechanisms, these are valuable observations that are simply not being reported.

The need for common standards was raised in relation to core aspects of working with LMM analysis, including model building, model comparison, model selection, and the interpretation of results. There are varying ways to build any statistical model, for example, in linear regression (e.g., stepwise model selection, simultaneous entry of covariate predictors) and so there are varying ways to build an LMM. There is no one approach that will suit all circumstances, therefore researchers should report and justify the process they took. A number of recent studies have shown how the results for experimental data can vary substantially depending on alternate more-or-less reasonable-seeming decisions taken during data analysis (Gelman & Loken, 2013; Silberzahn & Uhlmann, 2015; Simmons et al., 2011; see, also, Patel, Burford, & Ioannidis, 2015). The more complex the analysis pipeline, the greater the possible number of analyses, and the greater the likelihood of widespread but undocumented variation in practice. We do not identify the existence of alternate analytic pathways as inherently troubling – the path we take during analysis is always one amongst many. The difficulty for scientific reasoning stems from the occlusion of approaches, decisions and model features by inconsistent or incomplete reporting.

In general, maybe we as a field can live with a balance in which data are sacred but analyses are contingent. On publication, we share the data and analysis as transparently as possible, and seek to guarantee its fidelity. We do not assume that an analysis as-published will be the last word on the estimation of effects carried in the data. We allow that alternate analyses may, in future, lead to revision in estimated effects. This approach would be supported by a reduced reliance on significance cut-offs and a greater focus on effect sizes themselves. A more systematic exploration of the sensitivity of results to analytic choices may permit the field to build in robustness to results reporting. In a helpful recent discussion, Gelman and Hennig (2017) explore the ways in which researchers can usefully move to considering statistical analyses in terms of transparency, consensus, impartiality,

correspondence to observable reality, and stability. Consistent with our analysis of the application and reporting of mixed-effects models in psychological science, Gelman and Hennig (2017) advocate, moreover, the broader acknowledgement of multiple perspectives, the ways in which different decisions can be made given differing perspectives or in different contexts, and the rigorous explanations of our choices given the possibility of alternate approaches. It may be that we shall see, increasingly, that analyses addressing scientific hypotheses are supplemented by examinations of the stability of estimates over reasonable variants in approach. We are certain, however, that transparency in reporting will be foundational to progress.

4.1 Best Practice Guidance

In the following sections, we present short discussions and recommendations for practice for the key areas highlighted by the survey and review results. We offer, in Table 7, advice concerning best practice in reporting LMMs.

4.1.1 Preparation for using LMMs

A number of researchers are moving from analysing factorial design data with ANOVA to analysing factorial design data with LMMs. In this context, the sample of experimental stimuli or trial types needs to be carefully considered to furnish the sensitivity sufficient to detect experimental or observed effects (see below), and the computational engine (most often, maximum likelihood estimation) for LMMs assumes a large sample size (Maas & Hox, 2004; 2005). It is our view that some issues with convergence are likely caused by researchers using LMMs to analyse relatively small sets of data. With smaller samples, it is less likely that a viable solution can be found to fit the proposed model to the data. It is worth highlighting that the literature on mixed-effects models defines ‘small’ as 50 or fewer sampling

units (Bell et al., 2010; Maas & Hox, 2004; 2005). A researcher may be interested in the effect of frequency, testing this with 10 high frequency and 10 low frequency words. In an ANOVA, the participant average RT for the high and low frequency words would be calculated. In an LMM, this would be the coefficient for frequency (e.g. Figure 4a). However, a random effect may also be fit to model how this effect *differs for each participant* (e.g. Figure 4b). In this case, the model only has available 20 data points per participant (10 high and 10 low) and this may simply be insufficient to complete the computation (Bates et al., 2015). With more complex random effect structures (e.g., maximal structures for some designs, after Barr et al., 2013) and perhaps no change in how researchers plan experiments, it is unsurprising that convergence issues have become increasingly common.

In short, plan to collect data for as many stimuli and as many participants as possible. This comes with the caveat that with very large sample sizes, smaller effects can become 'significant' even though they may not be meaningful. We direct researchers to the discussion in Section 4.1.6 below, and the very sensible advice from the American Statistical Association (Wasserstein & Lazar, 2016) to move away from cut-offs for interpreting p values. Where smaller sample sizes are unavoidable (e.g. recruitment of hard to reach or specialist populations, difficulty generating large samples of stimuli), researchers should - of course - acknowledge this limitation. They should also examine (*see* osf.io/bfq39/files/ LMMs_BestPractice_Example_withOutput) the random effects and consider their validity. Convergence issues may mean that the fitting of random effects for some terms is not possible. Random effect variances that are close to zero indicate there is little variance to be accounted for in the data. Random intercepts and slopes that show high or near perfect correlations may indicate over-fitting.

4.1.2 Power Analysis for LMMs

It will surprise no-one that power analysis for LMMs is complicated. This is principally because study design features like the use of repeated measures require multiple level sampling (e.g., of participants, of stimuli) and entail a hierarchical or multilevel structure in the data (grouping trial-level observations, say, under participants or stimuli) (Scherbaum & Ferreter, 2009; Snijders & Bosker, 1993; Snijders, 2005). If, for example, a researcher presents all 20 stimuli to each of 20 participants, in each condition of a factorial design, the data sample can be characterized in terms of the lowest level of sampling (the individual observations, n=400, of each response by a participant to a stimulus) but also in terms of higher-level groupings, or sampling units (the number of participants and the number of items), while the mixed-effects model may incorporate terms to estimate effects or interactions between effects within and across levels of the hierarchical data structure (i.e. effects due to participant attributes, stimulus properties, or trial conditions). In addition, for LMMs, we can usefully consider the power to accurately estimate fixed effect coefficients, random effect variances, averages for particular sampling units or interactions across those units (Scherbaum & Ferreter, 2009; Snijders, 2005). From hereon we will focus only on power to detect fixed effect predictors.

For fixed effects, power in LMMs does not increase simply as the total sample of observations increases. Observed outcome values within a grouping (e.g., trial response values for a given participant) may be more or less correlated. If this correlation (the intra-class correlation for a given grouping) is high, adding more individual data points for a grouping does not add more information (Scherbaum & Ferreter, 2009). In other words, if the responses across trials from a particular participant are highly correlated, the stronger explanatory factor is the participant, not the individual trials or conditions (as we saw in the example in Section 1.1). Getting the participant to do more trials does not increase power. This also means that accurate power estimation for LMMs requires us to estimate or know the variation

within and between sampling units, e.g., for trials within subjects (Snijders & Bosker, 1993; Scherbaum & Ferreter, 2009). This is one of the reasons why reporting random effect variances is so important for the field.

The general recommendation is to have as many *sampling units* as possible, since this is the main limitation on power (Snijders, 2005), where sampling units consist of the sets by which the lowest level of observations (e.g., trial-level observations) are grouped, where groupings can be expected to cause correlations in the data (Bell, Morgan, Kromery & Ferron, 2010; Maas & Hox, 2005; 2006). Fewer sampling units will mean that effects estimates are less reliable (underestimated standard errors, greater uncertainty over estimates; Bell et al, 2010; Maas & Hox, 2004; 2005). When looking across a range of simulation studies, Scherbaum & Ferreter (2009) concluded that increasing numbers of sampling units is the best way to improve power (this held for the accuracy of estimating fixed effect coefficients, random effect variances and cross-level interactions). For psychological research, this means 30-50 participants, and 30-50 items or trials for each of those participants completing each condition (i.e. a total sample of 900-2500 data points; Scherbaum & Ferreter, 2009). For example, assuming typical effect sizes of 0.3-0.4 (scaled in standard deviations), Brysbaert and Stevens (2018) recommend a minimum of 40 participants and 40 items (1600 data points). It bears repeating that any power analysis is dependent on the effect sizes under consideration so there is no simple rule (e.g., “just use 40 participants and 40 items”). In parallel, it is an empty critique to say that studies are ‘underpowered’ unless we can guess the likely effect sizes. For example, with LMM analysis a typical factorial experiment in psychology with 30 participants responding to 30 stimuli has power of 0.25 for a small effect size (0.2) and 0.8 for a medium effect size (0.5, see Figure 2 in Westfall, Kenny & Judd, 2014). To achieve a power of 0.95 for this number of participants and stimuli, you need a minimum effect size of around 0.6. Recall that 0.4 is a typical effect size for psychological studies (Brysbaert & Stevens, 2018). Adding more participants alone

does not remedy this problem (Luke, 2016), as power asymptotes due to the variation in stimuli (Westfall, Kenny & Judd, 2014). This links back to the issue identified above: the higher-level groupings (sampling units) in the data influence variation (responses for the same participant are correlated, responses for the same items are correlated) so ideally, the numbers for all sampling units should be increased. Ultimately, these considerations *may change the design of the study*.

Brysbaert and Stevens (2018) provide an easy to read tutorial on conducting power analysis to detect fixed effects. They show how to use the online application from Westfall, Kenny and Judd (2014, jakewestfall.shinyapps.io/two_factor_power/) as well as power analysis using simulated data in R. For the online application from Westfall et al (2014), researchers need (a) an estimate of the effect size for the fixed effect (b) estimates for the variance components – i.e. the proportion of the total variance that each random effect in the model accounts for and (c) the number of participants and the number of items. For power analysis from simulation, researchers would ideally use pilot data or data from a published study. It is also possible, with some skill, to generate data sets that give an 'idealised' experiment outcome (e.g. a significant effect of some reasonable size) and base power analysis on that. It is worth stressing that without the full reporting of random effects in publications and more common sharing of data we are severely limiting our ability to conduct useful a-priori power analysis. Appendix 2 lists packages available for LMM power analyses, but we strongly recommend Brysbaert & Stevens (2018) as a starting point.

4.1.3 Assumptions for LMMs

Researchers should check whether the assumptions of LMMs have been met. For LMMs, we take the same assumptions as for regression (linearity, random distribution of residuals, homoscedasticy; Maas & Hox, 2004; 2005) except that LMMs are used because the independence assumption is violated because we know

that data are grouped in some way, so observations from those groups are correlated. For LMMs, we assume that residual errors and random effects deviations are normally distributed (Crawley, 2012: Field & Wright, 2011; Pinheiro & Bates, 2000; Snijders & Bosker, 2011). The simplest way to check these assumptions is to plot residuals and plot random effects. The script associated with Section 1.1 provides some R code for plotting random effects. For plotting residuals and checking model assumptions, we refer readers to the excellent tutorial by Winter (2013). It has been shown that non-normally distributed random effects do not substantially affect the estimation of fixed effect coefficients but do affect the reliability of the variance estimates for the random effects themselves (Maas & Hox, 2004).

4.1.4 Selecting Random Effects

The literature suggests that two approaches can sensibly be taken. Researchers may choose to select random effects according to experimental design (Brauer & Curtin, 2018; Barr et al 2013), and this can result in a maximal to minimal-that-converges modelling process (more on this below). Alternatively, researchers can select random effects that improve model fit (Bates et al., 2015; Linck & Cunnings, 2015; Magezi, 2015). This results in a minimal to maximal-that-improves-fit process. In both cases, the random effects part of the model is built first. Once it is established, fixed effects are added.

Selecting random effects according to experimental design has been recommended for confirmatory hypothesis testing (Barr et al, 2013) and this is the most common situation for researchers in experimental psychology. The steps are to identify the maximal random effects structure that is possible for the design, and then to see if this model converges (whether the model can be fit to the data). Brauer and Curtin (2018) helpfully summarise Barr et al (2013) with three rules for selecting a maximal random effects structure, add: (1) random intercepts for any unit (e.g.,

subjects or items) that cause non-independence in the data; (2) random slopes for any within-subject effects; and (3) random slopes for interactions that are completely within-subjects.

Many readers will have found that complex random effect structures may prevent the model from converging. This often occurs because the random effects specified in the model are not present in the data (Bates et al., 2015; Matuschek et al., 2017). For example, when a random effect is included to estimate variance associated with differences between participants in the effect of a within-subjects interaction between variables, while *in the data* the interaction does not substantially vary between participants, researchers would commonly find that the random effect cannot be estimated. Solutions to convergence problems may include the simplification of model structure (Brauer and Curtin, 2018; Matuscheck et al, 2017), using Principal Components Analysis to determine the most meaningful slopes (Bates et al., 2015), switching to alternate optimization algorithms (see comments by Bolker, 2015), or indeed to alternate programming languages or approaches (e.g., Bayes estimation, Eager & Roy, 2017). We strongly recommend the summary provided by Brauer and Curtin (2018), where a step-by-step guide is provided for dealing with convergence issues and, in particular, steps to take for simplification from a maximal model.

Alternatively, researchers may select random effects that improve model fit (Linck & Cunnings, 2015; Magezi, 2015). Matuscheck et al. (2017) demonstrated that models are more sensitive (in the detection of fixed effects) if random effects are specified according to whether Likelihood Ratio Test (LRT) model comparisons warrant their inclusion, that is, according to whether or not the random effects improve model fit. Matuschek et al. (2017) contend that we cannot know in advance whether a random effect structure is *supported* by the data, and that in the long run, fitting models with random effects selected for better model fit means that the researcher can effectively manage both Type I and Type II error rates. So, under this

process, the random effects are built up successively and tested at each point to see if they improve model fit, beginning with intercepts, slopes for main effects, then intercepts and slopes, and then interactions between main effects. Researchers may find that certain random effect terms do not improve model fit, or that the model does not converge with some terms. In the model output, random effect variances may be estimated as close to zero. Either outcome suggests the random effect being modelled is not present in the sample. Where covariances are modelled (correlations between intercepts and slopes), perfect correlations between random effect terms can indicate over-fitting. That is, all the variance explained by the random slope is in fact already explained by fitting the random intercept (leading to a perfect correlation between these terms). In this case, it is unlikely that the inclusion of the slope would improve model fit.

Our focus on random effects reflects the novelty of this requirement for psychological research, and the conceptual and computational challenges involved: what effects can be specified? (Barr et al., 2013); what effects allow a model to converge? (Eager & Roy, 2017). More broadly however, our discussion reflects a general point about model specification and selection: why should we want to build all models in the same way? The two options we have outlined above for selecting random effects are both reasonable and well-motivated. It should be left up to individual researchers to choose the approach they prefer and to give the rationale for that choice.

4.1.5 Model comparison and model selection

There is a tradition of data analysis in psychological research in which factorial ANOVAs are used to test all possible main effects and interactions, given a study design, in an approach that appears objective. We acknowledge that this approach appears to relieve the researcher of the need to make decisions about the model (though it may require decisions about the data, Steegen et al., 2016; and though

decisions may be involved in subtle ways, Gelman & Hennig, 2017; Simmons et al., 2011). It is tempting, therefore, for researchers to adhere to a prominent set of recommendations as the 'one true way' to complete analysis, disregarding the fact that LMMs require an explicit *modelling* approach. Comparable with other *modelling* approaches (e.g., growth curve modelling, structural equation modelling), however, we advocate that there should be a clear statement of the criteria used when selecting model parameters and these should be principally driven by the research questions.

4.1.5.1 A pragmatic approach to life with multiple models

It would be productive for the field if we acknowledge that the approach we take during analysis is typically to choose one course given alternatives. We should ask the questions "How was your study designed?" and "What do you want to know from the data?" and "Given that, why have you taken the approach you have taken?". So, it is inevitable that researchers will end up building and testing multiple models when working with LMMs. In the context of testing data from an experimental design (e.g., the kind of factorial design that would traditionally be analysed using an ANOVA), it is sensible for the fixed effects to be defined around the experimental conditions (see, e.g., Barr et al, 2013; Schad, Vasishth, Hohenstein & Kliegl, 2018). However, researchers may have fixed effect variables that they wish to analyse in addition to the experimental conditions. These could be added after the experimental conditions, added at the same time, or tested for inclusion. Naturally, the approach taken will depend on the hypotheses. As we have stated above, there is no single correct approach that will apply across all situations.

There are several approaches to model selection. In a controlled experimental study, the hypotheses about the fixed effects may be entirely specified in terms of the expected impact of the experimental conditions, and these could then be entered all at once (as for ANOVA). Alternatively, researchers may be interested in finding the

simplest explanation for the data. In this case, they might start with the most complex model, incorporating all possible effects implied by the experimental design, and remove terms that do not influence model fit (i.e., where a simpler model may explain the data comparably to a more complex model). The approach taken by a researcher should be justified with respect to their research questions, and assumptions.

4.1.5.2 Model comparison

Model comparison can be completed using information criteria (e.g., the Akaike Information Criterion, AIC, and the Bayesian Information Criterion, BIC; see discussions in Aho, Derryberry, & Peterson, 2014) and Likelihood Ratio Tests (LRTs). LRTs apply when models are *nested* (the terms of the simpler model appear in the more complex model) and the models are compared in a pairwise fashion (see discussions in Luke, 2016; Matuschek et al., 2017). If not nested, models can be evaluated by reference to information criteria. Aho et al. (2014) argue that AIC and BIC may be differently favoured in different inferential contexts (e.g., in their account, whether analyses are exploratory (AIC) or confirmatory (BIC)), and we highlight, for interested readers, a rich literature surrounding their use (e.g., Burnham & Anderson, 2004; see, also, McElreath, 2015). However, LRT model comparisons are often useful as a simple means to evaluate the relative utility of models differing in discrete components (models varying in the presence vs. absence of hypothesized effects). The LRT statistic is formed as twice the log of the ratio of the likelihood of the more complex (larger) model divided by the likelihood of the less complex (smaller) model (Pinheiro & Bates, 2000). It can be understood as a comparison of the strength of the evidence, given the same data, for the more complex versus the simpler model. The likelihood comparison yields a p-value (e.g., using the anova() function in R) because the LRT statistic has an approximately χ^2 distribution, assuming the null hypothesis is true (that the simpler model is adequate), with degrees of freedom equal to the difference between the models in the number of terms.

When comparing models using LRTs, successive models should differ in either their fixed effects or their random effects but not both. This is because (a) models tested with LRTs must be nested and (b) a change in the random effect structure will change the values of the fixed effects (and vice versa). Models can be generated using maximum likelihood (ML) or restricted maximum likelihood (REML). Both methods solve model fitting by maximizing the likelihood of the data given the model. When comparing models that differ in their fixed effects, it is recommended to use ML estimation for the models. This is because REML likelihood values depend on the fixed effects in the model (Faraway, 2016; Zuur et al, 2009). When comparing models that differ in their random effects, it is recommended to use REML estimation for the models. This is because ML estimates of random variance components tend to be underestimated in comparison with REML estimates (Zuur et al, 2009).

Researchers may be concerned whether there need to be corrections for multiple comparisons when multiple models are being compared using LRTs. If a complex model is being built and LRTs are being used at each step to judge the inclusion or exclusion of a particular effect, should there be an adjustment to the alpha level to reflect the volume of comparisons being made? The problem can be framed in terms of the simplification of a model where greater complexity is rejected because the more complex model is found, by means of the LRT comparison, to fit the data no better than the simpler model. A simpler model, in that circumstance, will be associated with too narrow confidence limits and too small p-values, however good the overall fit, because degrees of freedom corresponding to the dismissed complexity (the rejected larger model) are then not accounted for in the estimation of standard errors for the simpler model (cf. Harrell, 2001). More generally, p-values depend upon the researcher following their intentions: adhering to prior sampling targets, or completing as many statistical comparisons as were planned (Kruschke, 2013). Therefore, our advice would be that, firstly, researchers should be explicit about the models they fit and evaluate. Secondarily, if researchers plan to perform

significance tests, they should consider the utility of pre-registering experimental data collection and analysis plans (Nosek, Ebersole, DeHaven, & Mellor, 2018).

4.1.5.3 Using multiple models to test for robust effects

It is worth considering how variation in data preparation and model building can be harnessed to clarify the stability or sensitivity of effect estimates. Steegen et al. (2016) described multiverse analyses, in which all possible data sets are constructed from a sampling of the alternative ways in which raw data can be prepared for analysis (e.g., with variations on outlier exclusion, variable coding) and the analysis of interest is then performed across these data sets. P-value plots can be used to show how effects vary across differently collated data sets, indicating the robustness of results, or potential holes in theory or measurement. Patel, Burford and Ioannidis (2015) describe the "vibration of effects" or VoE which shows the variation in effect estimates across different models. This is particularly applicable in cases where there are many ways to specify models, and many possible variables or covariates of interest. VoE analysis shows how the influence of a variable changes across models and as more covariates are included (adjustment variables).

4.1.5.4 Reporting Model building

The problem we have identified, the arbitrary variation in reporting and analytic practice, is *not* insoluble. When multiple models have been fit to reach a final 'best model', best practice is to report the process of comparison. Appendix Table A5.1 offers a format for reporting LRT model comparisons concisely. When multiple, equally plausible, models of the data are possible, a fruitful approach is to examine the variation in estimates across a series of models and report this as a test of the robustness of effects (Patel et al, 2015).

In an era of online publication, it is straightforward for appendices and supplementary materials to house additional information. The provision of analysis

scripts and data with publication are a straightforward means to repeat or modify analyses if researchers (and reviewers) so wish. With the increasing use of pre-registration, researchers will specify in advance the modelling approach they will use. This may include an actual model to be fit (i.e. a model equation), but at minimum it should include the dependent variable(s), fixed effects, covariates, a description of how random effects were chosen and the method by which model selection will take place (e.g. simple to complex, covariates first etc.). To be truly comprehensive it should also have an a-priori power analysis (see section 4.1.2); this alone would mean the model (or alternative models) are well specified beforehand.

4.1.6 Testing the significance of fixed effects

Researchers familiar with ANOVA will know that significance tests typically require the specification of model and error (denominator) degrees of freedom. Computing degrees of freedom for significance tests in LMMs is a non-trivial problem (Baayen at al., 2008; Bates, 2006; Luke, 2016). For models with a hierarchical structure it is not clear how to define the denominator degrees of freedom (e.g., by number of observations, number of participants, or number of random effects). As Luke (2016) notes, researchers may prefer to use model comparison with LRTs to evaluate the significance of a fixed effect as this method does not require computation of denominator degrees of freedom. The lme4 package in R (Bates et al, 2015) provides a summary guide to how p-values can be obtained for fitted models (search for help("pvalues") when lme4 is installed), with a number of different options for confidence intervals, model comparison and two named methods for computing degrees of freedom (Kenward-Roger, Satterthwaite). Clearly, one reason why multiple methods for computing p-values appear in the literature is that a variety of options are available.

Luke (2016) used simulations to compare different methods for computing significance in LMMs. In pairwise model comparisons, observed likelihood ratios are

associated with p-values under the assumption that the distribution of the LRT statistic approximates the χ^2 distribution. Alternatively, the t statistics associated with model coefficients can be treated as z scores, where t > 1.96 effects can be taken to be significant (at the .05 alpha level). Luke (2016) found that interpreting t as z is anti-conservative, especially for small samples of participants and items and, critically, that this risk is independent of the total number of observations because one cannot compensate for small numbers of participants with large numbers of items. In our literature review, LRTs and t-as-z approaches were the most commonly used in published manuscripts. Luke (2016) reports that Satterthwaite and Kenward-Rogers approximations when applied to models estimated with REML yield relatively robust significance tests across different samples sizes. Following Luke (2016), we recommend the use of these methods when p-values are needed for fixed effects. If researchers want to complete the equivalent of ANOVA omnibus and follow up tests, they can perform an LRT when a fixed effect is added to the model (omnibus test) and then compute contrasts (the follow up tests) from the model (see Schad, Vasishth, Hohenstein & Kliegl, 2018, for detailed guidance on performing contrasts in R). In summary, once the final model is established, it can be estimated with REML, and significance tests for model coefficients can be performed using Satterthwaite or Kenward-Rogers approximate degrees of freedom.

Alternatively, some researchers argue for abandoning dichotomous “above or below 0.05” thresholds (Amrhein, Greenland & McShane, 2019; Wasserstein, Schirm & Lazar, 2019; Wasserstein & Lazar, 2016). This is in line with a now substantial body of work arguing for a change in how Null Hypothesis Significance Testing (NHST) and frequentist statistics are used. For example, reporting means or coefficient estimates and confidence intervals but not p-values (Cumming, 2013a; 2013b) or interpreting p-values as just another piece of information about the likelihood of the result (Wasserstein, Schirm & Lazar, 2019). We strongly advise

readers to familiarize themselves with the American Statistical Association's statement on p-values (Wasserstein & Lazar, 2016).

An increasing number of researchers advocate the adoption of Bayesian analysis methods (Kruschke, 2013; McElreath, 2015) in which estimates for fixed effects coefficients and random effects variances (or covariances) are associated with posterior distributions that allocate varying probabilities to different potential effect values. Researchers familiar with lme4 model syntax (Baayen et al., 2008; Bates et al., 2015) can apply the same syntax to fit Bayesian mixed-effects models (using the brms library, Burkner, 2017). With Bayesian models, researchers can identify the credible interval encompassing the plausible estimates for an effect (see Vasishth, Nicenboim, Beckman, Li, & Kong, 2018, for a helpful recent tutorial; see Nicenboim & Vasishth, 2018, for an example report in this journal) instead of seeking to test (only) the existence of the effect (Kruschke, 2013). Bayesian model fitting encourages the incorporation of prior beliefs about the varying plausibility of potential estimates for target effects. For example, researchers interested in the effect of word attributes on response latency in reading tasks would, perhaps, suppose *a priori* that the coefficient for a hypothesized effect in this domain would be captured by an estimate associated with a normal probability distribution centered on 0, with a standard deviation of plus or minus 10. This quantifies the belief that psycholinguistic effects vary in size and direction, are of the order of tens of milliseconds, and that some hypothesized effects may tend to zero. Relevant to earlier discussion, recent work has shown that problems encountered with convergence for more complex mixed-effects models can be avoided through using Bayesian model fitting given the specification of prior information (Eager & Roy, 2017). Essentially, this is because the incorporation of prior information directs model fitting processes away from extreme values (e.g. random effects variances close to zero) that can cause problems for convergence. Regardless of whether models are fit with frequentist or Bayesian

methods, reporting of the modelling process needs to be entirely transparent.

4.1.7 Reporting

The standard for publication should be that other researchers can reproduce the study itself, as well as the study's results on the basis of the reported method, analysis approach and data (if available) (e.g., Open Science Collaboration, 2015). It is our judgment that many issues arise because of 'under-reporting' – that is, insufficient information provided in publications on the analysis steps (Gelman & Loken, 2013; Silberzahn & Uhlmann, 2015; Simmons et al., 2011) and for LMMs more specifically, incomplete reporting of model results. Table 7 provides guidance for the reporting of LMMs (more specific guidance on Generalised Linear Mixed-effects Models can be found in Bolker et al., 2009).

We have been asked what to do about the extensive documentation required by what we see as best practice, comprehensive, reporting. The simple solution is for researchers to share their data analysis scripts with publication. Scripts show exactly what decisions have been taken and exactly how models were selected and compared. When provided with data, they allow any other researcher to replicate entirely the reported results. Researchers using R may also consider making their whole analysis reproducible (Marwick, Boettiger, Mullen, 2018). This can be achieved with packages such as docker, which creates a container (a stand-alone application, Gallagher, 2017). This recreates the complete environment of the original analysis (for a tutorial, see Powell, 2019). The package holepunch will create a docker file, description and image on GitHub for a particular analysis that can then be run independently (Ram, 2019). For long term storage of scripts and analysis information there are a number of options where journal space is tight – many institutions provide data storage and archive facilities for their researchers, and the Open Science Framework provides facilities for data storage and archive, as well as pre-registration and project documentation.

Knowing in advance that an analysis script will be shared on publication will likely make researchers more systematic and attentive to their code and annotations in the first place. It should also encourage more supportive discussion (rather than criticism) around analysis processes and best practice methods, and give the less experienced an easy way to learn from experts.

Table 7: Best practice guidance for reporting LMMs

Issue	Recommendation
Preparation for modelling	
Software	Report the software and version of software used for modelling
Power analysis (section 4.1.2)	Report any a-priori power analyses, including effect sizes for fixed effects and variances for random effects.
The model	
Assumptions of LMM (section 4.1.3)	Report what data cleaning has been completed, outlier/data removal, transformations (e.g., centering or standardizing variables) or other changes prior to or following analysis (e.g., Baayen & Milin, 2015). Report whether models meet assumptions for LMMs. Report if transformations were carried out in order to meet assumptions (e.g., log transformation of reaction time to meet the assumption that residuals are normally distributed).
Selection of fixed and random effects (section 4.1.4 and 4.1.5)	Random effects are explicitly specified according to sampling units (e.g., participants, items), the data structure (e.g., repeated measures) and anticipated interactions between fixed effects and sampling units (e.g., intercepts

	only or intercepts and slopes). Fixed effects and covariates are specified from explicitly stated research questions and/or hypotheses. Report the size of the sample analysed in terms of total number of data points and of sampling units (e.g., number of participants, number of items, number of other groups specified as random effects, such as classes of children).
Model comparison* (section 4.1.5)	A clear statement of the methods by which models are compared/selected; e.g., simple to complex, covariates first, random effects first, fixed effects first etc. Report comparison method (LRT, AIC, BIC) and justify the choice. A complete report of all models compared (e.g., in appendices/supplementary data/analysis scripts) with model equations and the result of comparisons. An example table reporting model comparisons can be found in Appendix Table A5.1.
Convergence issues (section 4.1.5)	If models fail to converge, the approach taken to manage this should be comprehensively reported. This should include the formula for each model that did or did not converge and a rationale for a) the simplification method used and b) the final model reported. This may be most easily presented in an analysis script.
The results (section 4.1.6 and 4.1.7)	
Model*	Provide equation(s) that transparently define the reported model(s). An elegant way to do this is providing the model equation with the table that reports the model output (see

	Appendix Table A5.2).
Model output*	Final model(s) reported in a table that includes all parameter estimates for fixed effects (coefficients, standard errors and/or confidence intervals, associated test statistics and p-values if used), random effects (standard deviation and/or variance for each random effect, correlations/covariances if modelled) and some measure of model fit (e.g. R-squared, correlation between fitted values and data) (see Appendix Table A5.2).
Data and code	Share coding script used to complete the analysis. Wherever possible share data that generated the reported results.

* Example tables here are adapted from the excellent examples in Stevenson et al., 2013 (Table 2), Goldhammer et al., 2014 (Table 1) and Li et al., 2014.

5.0 Conclusion

We completed a survey of current practice and a review of published papers for LMMs. Concerns raised in the survey were broadly corroborated by data from a review of published papers. In response to this, we have reviewed current guidelines for the implementation and reporting of LMMs, and provided a summary of best practice. A summary of that summary is provided below. The survey highlighted that many researchers felt they had a lack of knowledge, or were unable to properly deal with the complexity of LMMs. We hope this paper has gone some way to remedying this deficit (perceived or real), and encouraging researchers to spend time preparing analyses in a such a way that fully transparent reporting is painless.

5.1 Bullet points for Best Practice

- Plan to collect data for as many stimuli and as many participants as possible.
- Complete power analysis prior to data collection. This will require that you specify the model and consider plausible effect sizes.
- Acknowledge that the choices you make during analysis are considered, justified and one path amongst many.
- During analysis, check that assumptions of LMMs have been met.
- If using LMMs to control for unexplained variance (e.g. when replacing ANOVA), fit random effects first.
- Provide a clear rationale for selection of fixed effects and any model comparison or model selection process.
- Appendix 5 provides example tables for concisely reporting model comparison and model outputs (https://osf.io/bfq39/files/)
- Provide the model equation(s) for the final model or models to be reported.
- If reporting p values, estimate the final model or models to be reported using REML and report Satterthwaite or Kenward-Rogers approximate degrees of freedom for p values for fixed effect coefficients.
- Report point estimates, standard errors and confidence intervals for the fixed effect coefficients.
- Report random effect variances from the final model in full.
- Whenever possible, share analysis code and data on publication.

6.0 References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience*, 17(4), 491-496.

Aho, K., Derryberry, DW & Peterson, T. (2014) Model selection for ecologists: the worldview of AIC and BIC, *Ecology*, 95(3), 631-636.

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567*, 305-307.*

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1), 149-157.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412.

Baayen, R.H. (2013). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.4.1. http://CRAN.R-project.org/package=languageR

Baayen, R. H., & Milin, P. (2015). Analyzing reaction times. International *Journal of Psychological Research*, 3(2), 12-28.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278.

Bates, D.M. (2006) [R] lmer, p-values and all that. Post on the R-help mailing list, May 19th, available at: https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D. M. (2007). Linear mixed model implementation in lme4. Manuscript, university of Wisconsin - Madison, January 2007.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv* preprint arXiv:1506.04967.

Bell, B.A., Morgan, G.B., Kromery, J.D. & Ferron, J.M. (2010) The Impact of Small Cluster Size on Multilevel Models: A Monte Carlo Examination of Two-Level Models

with Binary and Continuous Predictors. *JSM Proceedings, Survey Research Methods Section*, *1*(1), 4057-4067.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.

Boisgontier, M. P., & Cheval, B. (2016). The ANOVA to mixed model transition. *Neuroscience & Biobehavioral Reviews*, 68, 1004-1005.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24, 127-135.

Bolker, B. (2015, March 14). GLMM. Retrieved August 01, 2016, from http://glmm.wikidot.com/faq

Bowen, N. K., & Guo, S. (2011). *Structural equation modeling*. Oxford University Press.

Brauer, M., & Curtin, J. J. (2018). Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods*, *23*, 389-411.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. London, UK: Sage.

Brysbaert, M. (2007). *The language-as-fixed-effect-fallacy: Some simple SPSS solutions to a complex problem*. London: Royal Holloway, University of London.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1(1).

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1-28.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304.

Burstein, L., Miller, M.D., & Linn, R.L. (1981). *The Use of Within-Group Slopes as Indices of Group Outcomes*. Center for the Study of Evaluation, Graduate School of Education, UCLA, Los Angeles California. CSE Report 171.

Carp, J. (2012a). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1), 289-300.

Carp, J. (2012b). On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6, 149.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain

Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological Science*, 2515245919858072.

Chabris, C.F., Hebert, B.M., Benjamin, D.J., Beauchamp, J., Cesarini, D., van der Loos, M., Johannesson, M., Magnusson, P.K.E., Lichtenstein, P., Atwood, C.S., Freese, J., Hauser, T.S., Hauser, R.M., Christakis, N. & Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, 1(23), 1314-1323.

Chang, Y-H. & Lane, D.M. (2016) Generalizing across stimuli as well as subjects: A non-mathematical tutorial on mixed-effects models. *The Quantitative Methods for Psychology*, 12 (3), 201-219. 10.20982/tqmp.12.3.p201.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335-359.

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement* 7(3), 249-253.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple correlation/regression analysis for the behavioral sciences*. UK: Taylor & Francis.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, 14(1), 219-226.

Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

Cumming, G. (2013a). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2013b). The new statistics why and how. *Psychological Science*, *25*(1), 7-29.

Cunnings, I. (2012) An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28 (3), 369-382.

Davies, R.A.I., Arnell, R., Birchenough, J., Grimmond, D., & Houlson, S. (2017). Reading Through the Life Span: Individual Differences in Psycholinguistic Effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1298-1338.

Eager, C., & Roy, J. (2017). Mixed effects models are sometimes terrible. *arXiv* preprint arXiv:1701.04858.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Field, A., Miles, Z. & Field, Z. (2009). *Discovering statistics using R*. London, UK: Sage publications.

Field, A. & Wright, D.B. (2011) A Primer on Using Multilevel Models in Clinical and Experimental Psychopathology Research. *Journal of Experimental Psychopathology*, 2(2), 271-293.

Frisson, S., Koole, H., Hughes, L., Olson, A., & Wheeldon, L. (2014). Competition between orthographically and phonologically similar words during sentence reading: Evidence from eye movements. *Journal of Memory and Language*, 73, 148-173.

Gallagher, S. (2017). *Mastering Docker*. Packt Publishing, USA.

Gelman, A. (2014). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management*, 41 (2), 632-643. DOI: 0149206314525208.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 967-1033.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Model*s. Cambridge, UK: Cambridge University Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608-626.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

IBM Corp (2013 release). *IBM SPSS Statistics for Windows*, Version 22.0. Armonk, NY: IBM Corp.

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, 18(4), 40-47.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434 – 446.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, 15(2), 281-320.

JASP Team (2016). *JASP* (Version 0.8.0.0) [Computer software].

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135, 12-35.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1, 238.

Kliegl, R. (2014). Reduction of complexity of linear mixed models with double-bar syntax. RPubs.com/Reinhold/22193

Kreft, I.G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5), 535-540.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573-603.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30. http://CRAN.R-project.org/package=lmerTest

Li, X., Bicknell, K., Liu, P., Wei, W., & Rayner, K. (2014). Reading is fundamentally similar across disparate writing systems: A systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General*, *143*(2), 895.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience*, 4(4), 423-428.

LimeSurvey Project Team & Schmitz, C. (2015) LimeSurvey: An Open Source survey tool /LimeSurvey Project Hamburg, Germany. URL http://www.limesurvey.org

Linck, J. A., & Cunnings, I. (2015). The Utility and Application of Mixed-Effects Models in Second Language Research. *Language Learning*, 65(S1), 185-207.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, 39(4), 723-730.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 149.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior research methods*, *49*, 1494-1502.

Maas, C.J.M. & Hox, J.J. (2004) The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, 46, 427-440.

Maas, C.J.M. & Hox, J.J. (2005) Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, 1(3), 86-92.

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, 6, https://doi.org/10.3389/fpsyg.2015.00002

Marwick, B., Boettiger, C., & Mullen, L. (2018) Packaging data analytical word reproducibly using R (and friends) PerrJ Preprints 6:e2192v2.

MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.

McCoach, B.D., Rifenbark, G.G., Newton, S.D., Xiaoran, L., Kooken, J., Yomtov, D., Gambino, A.J., & Bellara, A. (2018) Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, 43 (5), 594-627.

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Meteyard, L., & Bose, A. (2018). What does a cue do? comparing phonological and semantic cues for picture naming in aphasia. Journal of Speech, Language, and Hearing Research, 61(3), 658-674.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1287.

Nagin, D.S. & Odgers, C.L. (2010). Group-Based Trajectory Modeling in Clinical Research. *Annual Review of Clinical Psychology*, 6, 109-138.

Nava & Marius (2017, May 23). Glmer mixed models inconsistent between lme4 updates. Retrieved July 11, 2019, from https://stackoverflow.com/questions/20963216/glmer-mixed-models-inconsistent-between-lme4-updates

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1-34.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*), 2600-2606.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6), 528-530.

Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046-1058.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer-Verlag.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2016). nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-128, URL: http://CRAN.R-project.org/package=nlme.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17(11), 1510.

Powell, D. (2019). Conducting reproducible research with Docker (Part 1 of 3). Retrieved from http://www.derekmpowell.com/posts/2018/02/docker-tutorial-1/

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raaijmakers, J. G. (2003). A further look at the" language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 57(3), 141.

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, 41(3), 416-426.

Ram, K. (2019). Hole punch. Retrieved 15 August, 2019, from https://karthik.github.io/holepunch/index.html

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009). *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata (3rd Edition)*. College Station, TX: Stata Press.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. & Lewis, T. (2000). *A user's guide to MLwiN*. London: Institute of Education, 286.

Rietveld, T., & van Hout, R. (2007). Analysis of variance for repeated measures designs with word materials as a nested random or fixed factor. *Behavior Research Methods*, 39(4), 735-747.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. Aphasiology, 24, 121–133.

Rossini, A. J., Heiberger, R. M., Sparapani, R. A., Maechler, M., & Hornik, K. (2004). Emacs speaks statistics: A multiplatform, multipackage development environment for statistical analysis. *Journal of Computational and Graphical Statistics*, 13(1), 247-261. URL: https://ess.r-project.org/

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2018). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. arXiv preprint arXiv:1807.10451.

Scherbaum, C.A., & Ferreter, J.M. (2009) Estimating Statistical Power and Required Sample Sizes for Organisational Research Using Multilevel Modeling. *Organizational Research Methods*, 12(2), 347-367.

Schluter, D. (2015). Fit models to data. Retrieved August 1, 2016, from https://www.zoology.ubc.ca/~schluter/R/fit-model/

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature*, 526(7572), 189.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Snijders, T.A. (2005) Power and Sample Size in Multilevel Linear Models. In: B.S. Everitt and D.C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Volume 3, 1570–1573. Chices- ter (etc.): Wiley.

Snijders, T.A., & Bosker, R.J. (2011). *Multilevel analysis (2nd Edition).* London, UK: Sage.

Snijders, T.A., & Bosker, R.J. (1993) Standard Errors and Sample Sizes for Two-Level Research. *Journal of Educational Statistics*, 18(3), 237-259.

Stan Development Team (2016). *Stan Modeling Language Users Guide and Reference Manual*, Version 2.14.0. http://mc-stan.org

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702-712.

Stevenson, C. E., Hickendorff, M., Resing, W. C., Heiser, W. J., & de Boeck, P. A. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence*, *41*(3), 157-168.

Th. Gries, S. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1), 95-125.

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed - effects regression with R examples. *Psychophysiology*, 52(1), 124-139.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1), 126-156.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147-161.

Venables, W. N. 2014. *S-PLUS and S*. Wiley StatsRef: Statistics Reference Online.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3), 274-290.

Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2), 150-158.

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. New York, NY: Oxford University Press.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". Editorial. *The American Statistician*, 73 (Issue supplement 1: Statistical Inference in the 21st Century: A World Beyond p < 0.05), 1-19.

Wasserstein, R.L. & Lazar, N.A. (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133.

West, B. T., & Galecki, A. T. (2011) An overview of current software procedures for fitting linear mixed models. *The American Statistician*, 65(4), 274-282.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv* preprint arXiv:1308.5499.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3-36

Wood, S. N. & Scheipl, F. (2016). gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'. R package version 0.2-4.
http://CRAN.R-project.org/package=gamm4

Wright, D. B., & Villalba, D. K. (2012). Memory conformity affects inaccurate memories more than accurate memories. *Memory*, 20(3), 254-265.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed*

effects models and extensions in ecology with R. Springer Science & Business Media.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 386-397.

Titles and Legends for Figures 1-5

Figure 1 Title: Illustrations for Participant Intercepts for Naming Accuracy
Figure 1 Legend: Figure 1a shows each participant's mean accuracy across all the naming trials they completed, with the group mean as the rightmost column. Figure 1b shows the participant's accuracy scaled as standard deviations from the group mean (centered at zero) – the Random Intercepts by Participant.

Figure 2a Title: Average accuracy across the four Cue Type conditions
Figure 2a Legend: Accuracy values are fitted values taken from a mixed effect model fit to the data with all three fixed effect predictors and random slopes for Cue Type by Participant: Accuracy ~ Cue Type + Length Phonemes + Frequency + (0 + Cue Type | Participant). Note random intercepts for Participants were not included in this model, to illustrate a slopes only model. Error bars are 95% confidence intervals.

Figure 2b Title: Effect of Cue Type by Participant
Figure 2b Legend: Sh.Ons = Shared onset cue (phonological cue), Assoc = Associated word cue, NonAssoc = Non associated word Cue. Each panel represents the data from a single participant, showing their naming accuracy (across all trials in that Cue Type condition) as a boxplot. Accuracy values are fitted values taken from a mixed effect model fit to the data with all three fixed effect predictors and random slopes for Cue Type by Participant: Accuracy ~ Cue Type + Length Phonemes + Frequency + (0 + Cue Type | Participant). Note random intercepts for Participants were not included in this model, to illustrate a slopes only model.

Figure 2c Title: Effect of Cue Type by Participant, as deviations from the condition mean (Random slopes for Cue Type by Participant)
Figure 2c Legend: Each panel shows the values for a Cue Type condition (Shared onset, Tone, Associated word, Non Associated word). In each panel, participant's accuracy is scaled as standard deviations from the condition mean (centered at zero). These are the Random Slopes by Participant. These are taken from a mixed effect model fit to the data with all three fixed effect predictors and random slopes for Cue Type by Participant: Accuracy ~ Cue Type + Length Phonemes + Frequency + (0 + Cue Type | Participant). Note random intercepts for Participants were not included in this model, to illustrate a slopes only model.

Figure 3 Title: Illustrations for Participant Intercepts and Slopes for Length in Phonemes
Figure 3 Legend: Accuracy values are taken from a mixed effect model fit to the data with all three fixed effect predictors, random intercepts by Participant and correlated random slopes for Length by Participant: Accuracy ~ Cue Type + Length Phonemes + Frequency + (1 + Length Phonemes | Participant). Figure 3a shows the average effect of Length in Phonemes, a negative slope showing that words that are longer are harder to name. Figure 3b shows the effect of Length for each individual participant (steeper or shallower slopes) and the overall differences in accuracy between participants (higher or lower intercepts). Figure 3c shows the Random Intercepts and Slopes for Length. In Figure 3c the left panel shows the Participant

Intercepts, scaled as deviations from the grand mean Intercept (as in Figure 1b). The right panel of Figure 3c shows the Participant Slopes for the effect of Length scaled as deviations from the average effect of Length.

Figure 4 Title: Illustrations for Participant Intercepts and Slopes for Frequency.
Figure 4 Legend: These figures parallel those seen in Figure 3. Accuracy values are taken from a mixed effect model fit to the data with all three fixed effect predictors, random intercepts by Participant and correlated random slopes for Frequency by Participant: Accuracy ~ Cue Type + Length Phonemes + Frequency + (1 + Frequency | Participant). Figure 4a shows the average effect of Frequency, a positive slope showing that words with higher Frequency are easier to name. Figure 4b shows the effect of Frequency for each individual participant (steeper or shallower slopes) and the overall differences in accuracy between participants (higher or lower intercepts). Figure 4c shows the Random Intercepts and Slopes for Frequency. In Figure 4c the left panel shows the Participant Intercepts, scaled as deviations from the grand mean Intercept (as in Figure 1b). The right panel of Figure 3c shows the Participant Slopes for the effect of Frequency scaled as deviations from the average effect of Frequency.

Figure 5 Title: Number of Pubmed citations for ‘Linear Mixed Models’ by year
Figure 5 Legend: Generated using the tool available at http://dan.corlan.net/medline-trend.html, entering “Linear Mixed Models” as the phrase search term and using data from 2000 to 2018.